From THE INSTITUTE OF ENVIRONMENTAL MEDICINE
Karolinska Institutet, Stockholm, Sweden

GENE–ENVIRONMENT INTERACTIONS IN RHEUMATOID ARTHRITIS: QUANTIFICATION AND CHARACTERIZATION OF CONTRIBUTING FACTORS

Xia Jiang

姜 侠

Stockholm 2015

Published by Karolinska Institutet. Printed by US-AB

ISBN 978-91-7549-947-5

GENE–ENVIRONMENT INTERACTIONS IN RHEUMATOID ARTHRITIS: QUANTIFICATION AND CHARACTERIZATION OF CONTRIBUTING FACTORS

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Xia Jiang

Principal Supervisor:
Professor Lars Alfredsson
Karolinska Institutet
Department of Environmental Medicine
Division of Cardiovascular Epidemiology

Co-supervisor(s):
Professor Lars Klareskog
Karolinska University Hospital
Department of Medicine
Division of Rheumatology

Associate Professor Leonid Padyukov
Karolinska University Hospital
Department of Medicine
Division of Rheumatology

Assistant Professor Henrik Källberg
Karolinska Institutet
Department of Environmental Medicine
Division of Cardiovascular Epidemiology

Opponent:
Professor Alan Silman
University of Manchester
Department of Epidemiology
Division of Epidemiology

Examination Board:
Professor Claes-Göran Östensson
Karolinska Institutet
Department of Molecular Medicine and Surgery
Division of Pediatric Endocrinology

Associated Professor Karin Modig
Karolinska Institutet
Department of Environmental Medicine
Division of Epidemiology

Associate Professor Christopher Sjöwall
Linköping University
Department of Clinical and Experimental Medicine
Division of Rheumatology

To my grandfather Zidong Jiang

To my parents

致我的祖父姜子东

致我的父亲母亲

ABSTRACT

Rheumatoid arthritis (RA) is a chronic autoimmune inflammatory disease characterised by persistent synovitis, systemic inflammation and autoantibodies. RA has a complex aetiology with the involvement of genetic factors and environmental triggers, and their interactions. The inherited risk for RA is mostly attributed to multiple gene loci, of which the largest contribution is made by the major histocompatibility complex (*MHC*), also known as the human leukocyte antigen (*HLA*) genes, in particular linked to the *MHC* class II region. Shared epitope (SE) comprises a small part of the extensive MHC class II polymorphisms, and has been identified as the strongest genetic risk factor with each allele being associated with approximately a 2-fold increased RA risk. Cigarette smoking is the best-known environmental trigger and also increases RA risk approximately 2-fold. A profound SE–smoking interaction effect has been well described among different populations. The aim of the current thesis is to further characterise and quantify this gene–environment interaction, specifically: 1) to identify more gene–environment interaction signals using genome-wide materials; 2) to further explore the synergistic effect by identifying the interacting components (e.g. which chemical component in cigarette smoke triggers RA, the nicotine or the particles?); 3) to determine which amino acid positions of *MHC* class II loci interact with smoking, the traditional SE positions at HLA-DR or other regions such as HLA-B and HLA-DPB; and 4) to evaluate present understanding of the familial risk and heritability of RA, taking into account all the currently identified risk factors.

In Study I, we conducted a gene–smoking interaction analysis using the genetic information from the Immunochip and genome-wide association studies, in two separate Swedish case–control populations (the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study in Stockholm and a cohort from Umeå). We found no significant interaction signals outside of chromosome 6, in either anti-citrullinated protein/peptide antibody (ACPA)-positive or ACPA-negative RA, indicating that *HLA* remains a region of great importance, and well-powered studies with larger sample size are warranted to identify new signals.

In Study II, we performed association analysis between smokeless tobacco (snuff) and RA among EIRA subjects. We found that moist snuff use (current or past) was not related to the risk of either ACPA-positive or ACPA-negative RA. Analyses restricted to never smokers, or stratified by gender, provided similar results. We conclude that the use of moist snuff is not associated with the risk of either ACPA-positive or ACPA-negative RA, and the increased RA risk associated with smoking is therefore most probably not due to nicotine.

In Study III, we carried out interaction analysis between heavy smoking and RA-related amino acid positions (11, 13, 71, and 74 in HLA-DRβ1, 9 in HLA-B and 9 in HLA-DPβ1) using three separate case–control populations (EIRA, the Nurses' Health Study (NHS) and a Korean cohort). We found significant additive interactions between heavy smoking and the amino acid haplotype at HLA-DRβ1 in all populations. We further identified key interacting variants at HLA-DRβ1 amino acid positions 11 and 13, in addition to the traditional SE positions 71 and 74. Our findings suggest that a physical interaction between citrullinated auto-antigens produced by smoking and HLA-DR molecules is characterised by an HLA-DRβ1 four-amino acid haplotype, primarily by novel positions 11 and 13.

In Study IV, we determined to what extent familial risk of RA could be explained by established risk factors by linking EIRA subjects to nationwide registers. We found that established environmental risk factors did not explain the familial risk of either seropositive or seronegative RA to any

significant degree, and that currently known genetic risk factors accounted only for a limited proportion of the familial risk of seropositive RA. This suggests that many risk factors remain to be identified, in particular for seronegative RA. Therefore, family history remains an important clinical risk factor for RA.

LIST OF SCIENTIFIC PAPERS

I. An Immunochip-based interaction study of contrasting interaction effects with smoking in ACPA-positive versus ACPA-negative rheumatoid arthritis.
Xia Jiang,* Henrik Källberg,* Zuomei Chen, Lisbeth Ärlestig, Solbritt Rantapää-Dahlqvist, Sonia Davila, Lars Klareskog, Leonid Padyukov, Lars Alfredsson.
Rheumatology. Accepted.

II. Smokeless tobacco (moist snuff) use and the risk of developing rheumatoid arthritis: results from a case-control study.
Xia Jiang, Lars Alfredsson, Lars Klareskog, Camilla Bengtsson.
Arthritis Care & Research (Hoboken). 2014 Oct; 66(10):1582-6.

III. Interactions between amino-acid-defined MHC class II variants and smoking for seropositive rheumatoid arthritis.
Kwangwoo Kim,* Xia Jiang,* Jing Cui,* Bing Lu, Karen H. Costenbader, Jeffrey A. Sparks, So-Young Bang, Hye-Soon Lee, Yukinori Okada, Soumya Raychaudhuri, Lars Alfredsson, Sang-Cheol Bae, Lars Klareskog, Elizabeth W. Karlson.
Arthritis & Rheumatology. Accepted.

IV. To what extent is the familial risk of rheumatoid arthritis explained by established rheumatoid arthritis risk factors?
Xia Jiang,* Thomas Frisell,* Johan Askling, Elizabeth W. Karlson, Lars Klareskog, Lars Alfredsson, Henrik Källberg.
Arthritis & Rheumatology. 2015 Feb; 67(2):352-62.

*These authors contributed equally.

LIST OF OTHER RELATED SCIENTIFIC PAPERS

I. Anti-CarP antibodies in two large cohorts of patients with rheumatoid arthritis and their relationship to genetic risk factors, cigarette smoking and other autoantibodies.
Jiang X, Trouw LA, van Wesemael TJ, Shi J, Bengtsson C, Källberg H, Malmström V, Israelsson L, Hreggvidsdottir H, Verduijn W, Klareskog L, Alfredsson L, Huizinga TW, Toes RE, Lundberg K, van der Woude D. Ann Rheum Dis. 2014 Oct; 73(10):1761-8.

II. Improved performance of epidemiologic and genetic risk models for rheumatoid arthritis serologic phenotypes using family history.
Sparks JA,* Chen CY,* Jiang X, Askling J, Hiraki LT, Malspeis S, Klareskog L, Alfredsson L, Costenbader KH, Karlson EW. Ann Rheum Dis. 2014 Apr 30.

III. Genetic risk scores and number of autoantibodies in patients with rheumatoid arthritis.
Maehlen MT, Olsen IC, Andreassen BK, Viken MK, Jiang X, Alfredsson L, Källberg H, Brynedal B, Kurreeman F, Daha N, Toes R, Zhernakova A, Gutierrez-Achury J, de Bakker PI, Martin J, Teruel M, Gonzalez-Gay MA, Rodríguez-Rodríguez L, Balsa A, Uhlig T, Kvien TK, Lie BA. Ann Rheum Dis. 2015 Apr; 74(4):762-8.

IV. Polymorphisms in peptidylarginine deiminase associate with rheumatoid arthritis in diverse Asian populations: evidence from MyEIRA study and meta-analysis.
Too CL, Murad S, Dhaliwal JS, Larsson P, Jiang X, Ding B, Alfredsson L, Klareskog L, Padyukov L. Arthritis Res Ther. 2012 Nov; 14(6):R250.

V. Genetic and environmental determinants for disease risk in subsets of rheumatoid arthritis defined by the anticitrullinated protein/peptide antibody fine specificity profile.
Lundberg K, Bengtsson C, Kharlamova N, Reed E, Jiang X, Kallberg H, Pollak-Dorocic I, Israelsson L, Kessel C, Padyukov L, Holmdahl R, Alfredsson L, Klareskog L. Ann Rheum Dis. 2013 May; 72(5):652-8.

CONTENTS

LIST OF ABBREVIATIONS

A	Adenine
AA	Amino Acid
ACPA	Anti-citrullinated Protein/Peptide Antibody
AP	Attributable Proportion due to Interaction
BAL	Bronchoalveolar lavage
BMI	Body Mass Index
C	Cytosine
CI	Confidence Interval
CTLA4	Cytotoxic T-lymphocyte protein 4
DZ	Dizygotic twins
DMARD	Disease-Modifying Anti-Rheumatic Drug
EF	Excess Fraction
EIRA	Epidemiological Investigation of Rheumatoid Arthritis
FCRL3	Fc receptor-like protein 3
FDR	First-degree Relative
G	Guanine
GRS	Genetic Risk Score
GWAS	Genome-wide Association Study
HLA	Human Leukocyte Antigen
Ig	Immunoglobulin
IBD	Identical by Descent
LD	Linkage Disequilibrium
MI	Multiple Imputation
MHC	Major Histocompatibility Complex
MZ	Monozygotic twins
NARAC	North American Rheumatoid Arthritis Consortium
NSAID	Non-steroidal Anti-inflammatory Drug
OR	Odds Ratio
PADI4	Protein-arginine Deiminase type 4
PCA	Principal Component Approach

QC	Quality Control
RA	Rheumatoid Arthritis
RERI	Relative Excess Risk due to Interaction
RF	Rheumatoid Factor
RR	Relative Risk
S	Synergy Index
SE	Shared Epitope
SNP	Single-nucleotide Polymorphism
T	Thymine
TLR	Toll-like Receptor
TRAF1	Tumour Necrosis Factor Receptor Associated Factor 1

1 INTRODUCTION

Rheumatoid arthritis (RA), the most common inflammatory joint disease, occurs when the immune system mistakenly attacks its own tissue. RA is characterised by persistent synovitis, systemic inflammation and the presence of autoantibodies.[1-3] The disease affects 0.5–1% of the total population, in a female/male ratio of 2.5–3.0/1.[1-3] Estimates have shown that approximately 50% of the risk of developing RA is attributable to genetic constitution, of which the largest contribution is made by the major histocompatibility complex (*MHC*), also known as the human leukocyte antigen (*HLA*) loci, in particular the *MHC* class II (or *HLA-DR*) region. Shared epitope (SE), a small part of the HLA-DR, has been comprehensively investigated and identified as the RA genetic risk factor of primary impact.[4-6] The remaining disease susceptibility could be largely ascribed to environmental influences, of which smoking is the best-known risk factor, with a relative risk of between 1.5 and 2, and with a dose–response effect observed in several independent samples.[7-11] In addition to the genetic and the environmental factors, gene–gene and gene–environment interactions play important roles. A profound additive interaction between SE and smoking has been described and replicated in different populations.[12-14] Although all the above-mentioned results have been specifically restricted to anti-citrullinated protein/peptide antibody (ACPA)-positive RA, the underlying mechanisms remain to be elucidated, and much less is known about ACPA-negative RA.

Therefore, during the 4 years of study for this PhD, the gene–environment interactions in RA have been further explored in terms of the following questions:

- Firstly, from a genome-wide perspective, do interaction effects exist between smoking and genes outside of the *HLA* region?
- Secondly, which components of cigarette smoke contribute to this synergistic effect given that smoke is a complex mixture of chemical compounds, including nicotine, char and other adjuvants[15] which have different effects on the immune system?
- Thirdly, as understanding regarding the *HLA* region has increased considerably with in-depth analysis of amino acids (AAs) through imputation, which AA positions could be identified in this interaction effect? Is the interaction effect restricted to the traditonal SE positions at HLA-DR, or are other regions such as HLA-DP or/and HLA-B involved?
- Finally, current genome-wide association studies (GWAS) and large epidemiological investigations have identified a number of genetic and environmental risk factors in addition to smoking and SE; how much in total can these factors explain the RA heritability using familial aggregation as an indicator?

This thesis is based on four studies investigating the gene–smoking interaction as well as the proportion of heritability that can be explained by currently identified risk factors in RA. Epidemiological methods have been used to conduct these studies. It is hoped that this research may provide others with ideas for further novel research in this field.

2 BACKGROUND

2.1 IMMUNE SYSTEM, IMMUNITY AND AUTOIMMUNITY

The immune system consists of a collection of cells, tissues and molecules that mediate immunity, which is defined as resistance to disease, specifically infectious disease. The most important physiological function of the immune system is, through coordinated action against microbes, to prevent infections and to eradicate those that have become established. Moreover, the impact of the immune system goes beyond defence against infectious disease. On one hand, the immune system participates in the clearance of dead cells, even in some cases eradication of tumours, and in initiating tissue repair. In contrast to these beneficial roles, on the other hand, abnormal immune responses can injure cells and induce pathological inflammation, causing allergic, autoimmune and inflammatory diseases. In addition, the immune system recognises and responds to tissue grafts and newly introduced molecules, which provides a barrier to transplantation and gene therapy.

Host defence mechanisms consist of innate and adaptive immunity. Innate immunity is always present in healthy individuals, prepared to block the entry of microbes or rapidly eliminate those that enter host tissue. Therefore, the epithelial barriers of the skin usually provide the first line of defence in innate immunity. Phagocytes, natural killer cells and several plasma proteins, including the proteins of the complement system, attack microbes if they do penetrate the epithelium and enter tissues or the circulation. However, innate immunity is phylogenetically older than adaptive immunity, recognising structures shared by classes of microbes, and therefore has lower specificity. The defence against infectious microbes, especially those that are pathogenic in humans, requires a more specialised and powerful adaptive immune system, which consists of lymphocytes and their products (antibodies). Lymphocytes express receptors that specifically recognise a much wider variety of molecules produced by microbes as well as non-infectious substances (antigens). There are two types of adaptive immunity: humoral immunity mediated by antibodies that are produced by B cells mainly neutralises and eliminates microbes that are found outside host cells, and cell-mediated immunity mediated by T lymphocytes provides defence against microbes that live and divide inside infected cells. The innate and adaptive systems act in both separate and cooperative ways; for example, adaptive immune responses often involve cells of the innate immune system to eliminate microbes, but also enhance innate immunity.

There are several crucial properties of the adaptive immune system. It has a vast total population of lymphocytes consisting of many different clones, with each clone expressing an antigen receptor that is different from all others. This enables the immune system to respond to a vast number and variety of antigens, and also ensures that distinct antigens elicit responses that specifically target those antigens. Moreover, the adaptive immune system remembers the immune responses it has experienced and is capable of inducing more rapid, larger secondary immune responses to subsequent encounters with the same antigens. Finally, the immune system is able to react against an enormous number and variety of foreign

antigens, but it also has developed multiple regulatory systems to avoid re-activities against the host's own potentially antigenic substances, the so-called self antigens. This unresponsiveness to self, known as immunological tolerance, is the ability of the immune system to coexist with potentially antigenic self molecules, cells and tissues.

Immunological tolerance could be interpreted as a homeostatic process maintained by several existing mechanisms. Firstly because lymphocyte receptor specificities are generated in an unbiased way during the normal process of lymphocyte maturation, all types of receptor specificities will be generated irrespective of whether or not they possess the ability to recognise self antigen. The thymus (for T cells) and the bone marrow (for B cells) exert important functions in restricting the number of maturing self-reactive clones by negative selection mechanisms (central tolerance). Despite this negative selection, many self-reactive clones have already presented in the organisms. Therefore, several mechanisms must act in concert to prevent immune responses to self antigens. Indeed, when lymphocytes specific for self antigens encounter the particular antigens in the secondary lymphoid organs or peripheral tissues, a number of measures will be implemented to guarantee tolerance. As a result, these lymphocytes will undergo changes in their receptors (for B cells) or develop into regulatory cells, or anergy or apoptosis will be induced. Any errors or failures that occur during these processes will probably influence the maintenance of self-tolerance, and even lead to autoimmunity or autoimmune diseases. Multiple molecules have a role in the processes of maintaining central and peripheral tolerance, for example: autoimmune regulator is responsible for the thymic expression of many peripheral tissue antigens in central T lymphocyte tolerance; CD28, cytotoxic T lymphocyte-associated antigen 4 and programmed death protein 1 all function to terminate T cell activation, resulting in long-lasting T cell anergy; the transcription factor FoxP3 and cytokine transforming growth factor β are required for the development of regulatory T cells; and the binding of Fas and Fas ligand may induce programmed cell death of both T and B cells. The complexity of both the immune system and the maintenance of self-tolerance is a reflection of the genetic complexity of autoimmune diseases, to which multiple factors including the inheritance of susceptibility genes and environmental triggers contribute. Read the book written by Abul Abbas *et al.* for a detailed description of relevant concepts in basic immunolgy.[16]

2.2 RHEUMATOID ARTHRITIS

2.2.1 Clinical Features and Subclassification

Rheumatoid arthritis (RA) is a chronic, systemic, inflammatory autoimmune disorder characterised by progressive damage of synovial joints and variable extra-articular manifestations.[1-3] Disease onset is usually insidious with joint symptoms emerging over weeks to months and often accompanied by decreased appetite, weakness or fatigue; it can take several months before a firm diagnosis can be verified.[1-3] The major symptoms of RA are pain, stiffness and swelling of multiple peripheral joints, in a bilateral symmetrical pattern. The clinical course of the disorder can be extremely variable, ranging from mild, self-

limited arthritis or arthritis-related symptoms to rapidly progressive multisystem inflammation with severe morbidity and high mortality.[1-3]

The incidence of RA increases with age,[2] but the disease can occur at any age, most commonly affecting those aged 40–70 (mean 66) years.[17] RA is a common disease, estimated to affect about 0.5–1% of the total population worldwide, with a notable low prevalence in rural Africa and high prevalence among certain tribes of Native America.[2] The female/male ratio of RA is around 2.5–3.0/1.[18] In Sweden, data from both the Swedish National Patient Register and the Swedish Rheumatology Quality Register data have indicated a cumulative prevalence of RA of 0.77% (women 1.11%, men 0.43%) and an incidence of 41/100000 (women 56/100000, men 25/100000) up to 2008.[17,19] Moreover, the lifetime risk of RA among adults has been estimated to be 2.7% for women and 1.5% for men, meaning that 1 in 37 women and 1 in 67 men will develop RA during their lifetime.[17] As a disease in rapid transition, uncontrolled active RA causes joint destruction, functional disability, decreased quality of life and several comorbidities, which all account for early mortality.[20] The RA mortality rate has been continuously decreasing in recent decades, during which time treatment strategies have fundamentally changed, including an emphasis on early diagnosis and early intensive treatment, with the aim of slowing or preventing joint damage and remission as the major therapeutic goal.[21] Several medications have been introduced for treating RA: disease-modifying antirheumatic drugs (DMARDs), biologic agents, non-steroidal anti-inflammatory drugs (NSAIDs), corticosteroids, immunosuppressants and others.

It has been proposed that RA is best considered as a clinical syndrome spanning different disease subsets encompassing several inflammatory cascades, which all eventually leading towards a common pathway.[22] It has been increasingly recognised that dividing RA into at least two subgroups can account for the potentially different prevention and treatment strategies, as well as help to elucidate the distinct aetiology behind each subset.[1,3] The subdivision has been based firstly and classically on the presence or absence of rheumatoid factors (RFs), the key pathogenic markers (mainly Immunoglobulin (Ig) M and IgA RF) directed against IgG. Later, the classification has been based also on the presence or absence of ACPAs. Although in most cases, ACPA and RF status overlap among patients (i.e. ACPA-positive patients are more likely to be RF positive), ACPAs seem to be more specific for diagnosis and better predictors of poor prognosis. Therefore, in addition to the widespread classification criteria established in 1987 to define RA based on RF, a new set of criteria was developed in 2010 based on ACPAs, which will be discussed below.

2.2.2 Diagnostic Criteria

The diagnosis of RA remains criteria guided. The classification criteria that are currently well accepted and in widespread international use to define RA are the American College of Rheumatology (ACR) 1987 criteria (Table1),[23] which were derived by attempting to distinguish between patients with established RA and those with a combination of other definite rheumatologic diagnoses. They are therefore less helpful in identifying patients who

could benefit from early treatments, in other words those patients at a stage at which evolution of joint destruction can be prevented before initiation of a chronic erosive disease course.[24] The 2010 European League Against Rheumatism (EULAR)/ACR criteria for RA classification were subsequently developed (Table2), with the aim of facilitating the diagnosis of individuals at an earlier stage of disease.[24,25]

Table1. 1987 RA classification.[23]

Four of these seven criteria must be present. Criteria 1-4 must have present for at least 6 weeks.

Criterion	Definition
1. Morning stiffness	Morning stiffness in and around the joints, lasting at least 1 hour before maximal improvement
2. Arthritis of 3 or more joint areas	At least 3 joint areas simultaneously have had soft tissue swelling or fluid (not bony overgrowth alone) observed by a physician. The 14 possible areas are right or left PIP, MCP, wrist, elbow, knee, ankle, and MTP joints
3. Arthritis of hand joints	At least 1 area swollen (as defined above) in a wrist, MCP, or PIP joint
4. Symmetric arthritis	Simultaneous involvement of the same joint areas (as defined in 2) on both sides of the body (bilateral involvement of PIPs, MCPs, or MTPs is acceptable without absolute symmetry)
5. Rheumatoid nodules	Subcutaneous nodules, over bony prominences, or extensor surfaces, or in juxtaarticular regions, observed by a physician
6. Serum rheumatoid factor	Demonstration of abnormal amounts of serum rheumatoid factor by any method for which the result has been positive in <5% of normal control subjects
7. Radiographic changes	Radiographic changes typical of rheumatoid arthritis on posteroanterior hand and wrist radiographs, which must include erosions or unequivocal bony decalcification localised in or most marked adjacent to the involved joints (osteoarthritis changes alone do not qualify)

Table2. The 2010 ACR-EULAR classification criteria for RA.[24]

Target population (Who should be tested): Patients who

1. have at least 1 joint with definite clinical synovitis (swelling)
2. with the synovitis not better explained by another disease

Classification criteria for RA (score-based algorithm: add score of categories A - D, a score of ≥6/10 is needed for classification of a patient as having definite RA);

A. Joint involvement	Score
1 large joint	0
2-10 large joints	1
1-3 small joints (with or without involvement of large joints)	2
4-10 small joints (with or without involvement of large joints)	3
>10 joints (at least 1 small joint)	5
B. Serology (at least 1 test result is needed for classification)	
Negative RF and negative ACPA	0
Low-positive RF or low-positive ACPA	2
High-positive RF or high-positive ACPA	3
C. Acute-phase reactants (at least 1 test result is needed for classification)	
Normal CRP and normal ESR	0
Abnormal CRP or abnormal ESR	1
D. Duration of symptoms	
<6 weeks	0
≥6 weeks	1

2.2.3 Pathogenesis

In autoimmune diseases, the ability of the immune system to discriminate between self and non-self antigens fails. As a result, the individual's own cells and tissues are attacked by the immune responses, in RA targeting the synovium-lined small joints and subsequently involving other organs.[16]

A number of genes contribute to the development of autoimmunity in general, with particular linkages towards the *HLA* region. The association between *HLA* alleles and many autoimmune diseases has long been recognised and was one of the first indications that T cells played an important role in these disorders. Polymorphisms in non-*HLA* genes are also associated with various autoimmune diseases including, for example, protein tyrosine phosphatase N22 (*PTPN22*) in systemic lupus erythematosus and type I diabetes mellitus and nucleotide-binding oligomerisation domain-containing protein 2 (*NOD2*) in Crohn's disease.[26] Moreover, environmental triggers such as infection also predispose to autoimmunity, with possible mechanisms through inflammation and stimulation of expression of co-stimulators or cross-reaction between microbial and self antigens. One such example is rheumatic fever, which may occur following a bacterial throat infection.

Similar to the majority of autoimmune diseases, the aetiology of RA is not fully understood; however, it is clear that genetic constitution (with the primary risk factor being the SE on the *HLA* gene), environmental triggers (particularly smoking) and stochastic factors act in concert to cause this complex disease. Considerable research has defined several crucial cellular players in RA pathogenesis, including T cells, B cells, antigen-presentation cells, macrophages and others. Complex interactions among genes, environmental triggers, multiple immune cells, cytokines and proteinases mediate the disease.[27,28]

A potential model taking into account of all these factors has been proposed by Gary Firestein.[29] Briefly, in early RA, the activation of innate immunity probably occurs first. This serves as a key pathogenic mechanism for the initiation of synovial inflammation. Autoantibodies, such as RF and APCAs, engage with Fc receptors and represent an alternative mechanism of the inflammation initiation. Synovial dendritic cells activated by toll-like receptor (TLR) ligands can migrate to lymph nodes where activated T cells develop towards the T helper type 1 phenotype and, through chemokine receptors, migrate towards inflamed synovial tissue. After activation of innate immunity in the joints, the production of cytokines and expression of adhesion molecules allows the continued entrance of immune cells. In certain conditions, such as the presence of a suitable genetic background, lymphocytes may accumulate in inflamed synovium. Under these circumstances, the break in tolerance in connection with an *HLA-DR* background or the T cell repertoire might contribute to auto-reactivity towards newly exposed articular antigens. Eventually, long-standing disease could develop into a destructive form. Instead of a specific 'rheumatoid antigen', a wide variety of antigens can provide targets and cause both T cell activation and B cell maturation.

Hence, a combination of chance and pre-determined events and adaptive immune responses directed against autologous antigens are required for the progression of disease.

Karim Raza *et al.* proposed a system of six phases in the development of RA: genetic risk factors environmental risk factors, systemic autoimmunity, symptoms without clinical arthritis, unclassified arthritis and, finally, RA.[30] As shown in Figure1, the first two phases usually influence predisposition towards RA in a combined manner, followed by immune abnormalities, and no clinically apparent soft tissue swelling, then the first clinical features of synovitis until, eventually, the development of RA. Although it is often assumed that all individuals move sequentially through these phases, this might not necessarily be the case. Some patients might never experience all phases, some might pass through these phases in a different order, and some might even go backwards.[30,31] The individuals included in our study all had RA at phase F, and a majority of them (>85%) had symptom durations of less than 1 year.

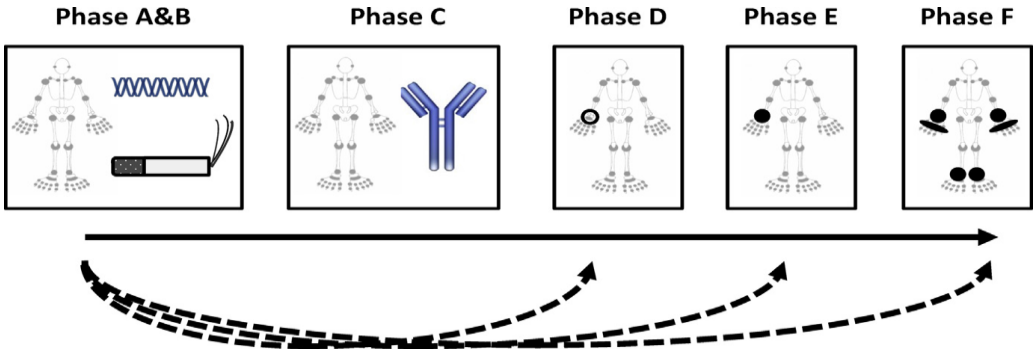


Figure1. The phases (A–E) that an individual may pass through in the transition from health to the development of RA (phase F). Adapted from Raza *et al.*.[30]

2.3 GENETICS IN RHEUMATOID ARTHRITIS

2.3.1 The Genome, Genes, Mutations and Polymorphisms

Genetics is the branch of science concerned with genes, heredity and variation in living organisms. The hereditary foundation of each living organism (e.g. bacteria, viruses and eukaryotes) is its genome, a long sequence of DNA that contains a complete set of hereditary information.[32] The human genome can be structurally divided into 22 autosomal chromosomes, the X and Y sex chromosomes and the extra-nuclear mitochondrial genome;[32] however, functionally, it is composed of genes, a sequence within the genome that gives rise to a discrete product such as a polypeptide or RNA. Each gene is a unit of a single stretch of DNA. DNA forms a double helix consisting of antiparallel strands where the nucleotide units are connected by 5' to 3' phosphodiester bonds, with the backbone on the exterior, and purine and pyrimidine bases are stacked in pairs in the interior via hydrogen bonds. Adenine (A) is complementary to thymine (T), and guanine (G) is complementary to cytosine (C). The human genome contains a total of ~3.3×10^9 base pairs of DNA, where only ~25% of the sequences are involved in producing proteins; among them, ~24% are introns (usually removed by subsequent RNA splicing) and only a tiny proportion (~1%) is accounted for by

the exons that actually code for polypeptides.[32] This reflects the large degree of DNA with unidentified function involving the human genome, which mainly consists of intergenic DNA (~22%) and repetitive sequences (~50%), with the latter being further composed of transposons, processed pseudogenes, simple sequence repeats, segmental duplications and tandem repeats.[32] Although their functions are generally unclear, the high proportion of the genome occupied by these elements might indicate active roles in shaping the genome. The number of valid genes is currently estimated to be in between 20000 and 25000, much less than originally expected.[32]

Individual genomes show extensive variations, and all genetic variance originates as mutations. Mutations are changes in the sequence of DNA, which can occur spontaneously or can be induced by mutagens.[32] Almost all organisms experience a certain extent of spontaneous mutation as the result of random interactions with the environment. Using mutagens, it becomes possible to induce numerous changes in any genes thus increasing the natural incidence of mutation (i.e induced mutagens). Mutations occur at multiple levels: across the whole genome, within a gene, or at a specific nucleotide site. A point mutation is the smallest mutation and changes only a single base pair. It can be caused by either chemical modification of DNA directly changing one base into another, or by errors during the replication of DNA through inserting the wrong base into a polynucleotide. A second common form of mutation is known as indels and comprises insertions or/and deletions. Indels of one or two base pairs can have the greatest effect if they are within the crucial coding sequences, due an inevitable frame-shift. Moreover, indels can affect parts of or even whole groups of genes. A possible source of mutations is the many different types of transposable elements, which are small DNA entities with mechanisms that enable them to move around and insert themselves into new locations. Mutational effects can be beneficial, harmful or neutral and can be reversible, depending on their context or location. In general, the more base pairs that are involved, the larger the effect of the mutation. Rather than a single mutation with great effect, most evolutionary changes are based on the accumulation of large numbers of mutations with small effects. Mutation is the main cause of diversity and once a mutation is carried with a frequency of more than 1% in the population, it is commonly known as a polymorphism. In this study, genetic measurements have been made mainly for the identification of single-nucleotide polymorphisms (SNPs), in which a single nucleotide (A, T, C or G) in the genome differs among the study population. Of note, only mutations in gametes can be transferred to the next generation, and many somatic mutations are not inherited.

Genetic predisposition (also known as genetic susceptibility) describes an increased likelihood of developing a particular disease or trait based on a person's genetic components, resulting from specific genetic variations that are inherited from parents and that contribute to the development of a disease with large or small effects. In a small minority of cases, genetic disorders can be caused by a single defective gene. Huntington's chorea, polycystic kidney disease, cystic fibrosis, phenylketonuria and haemophillia are typical examples of such disorders (*monogenic diseases*). They are usually inherited according to Mendel's laws as

being autosomal dominant, autosomal recessive or X-linked recessive.[33] In most cases, as for many common diseases, such as diabetes mellitus, schizophrenia and hypertension, strong genetic components are essential for their occurrence, which means that a large number of genes each functioning in a small but significant manner are needed to predispose individuals to these outcomes (*polygenic diseases*).[34,35] RA is a prototypical multifactorial trait, which is caused by the impact of various genes, each influencing the final outcomes to a small extent, as well as by interactions between multiple genes and often multiple environmental factors.

2.3.2 *MHC*, *HLA-DBR1* Gene, SE Hypothesis and RA

The human *MHC* is an unusual part of the genome, harbouring the highest density of genes (polygenic) with the majority exerting fundamental roles in immunity and having extremely high levels of variation (polymorphism) and extensive linkage disequilibrium (LD). The MHC was first demonstrated in mice and designated H (histocompatibility)-2 by geneticist George Snell, who proposed the idea of using congenic mice (i.e. mice that are bred to be genetically identical except at a single locus or genetic region) for the study of cancer. Snell quickly discovered that the genetic locus H-2 principally determined the status of acceptance or rejection for tumour grafts.[36] H-2 was subsequently shown to be a complex of many closely linked genes with many different alleles occurring at each locus, and was later termed the MHC. In the 1950s, Jean Dausset found iso-antibodies against leukocyte antigens in blood transfusion recipients, demonstrating a complex genetic system in humans similar to the H-2 system of mice.[37] He showed for the first time that the survival of a grafted kidney was correlated with the number of incompatibilities in the HLA system, which means the more similar the individuals are at their *HLA* locus, the more likely it is that they will accept grafts from one another. In addition to its immediate application to tissue transplantation, we now also know that of all regions identified so far, the *MHC* region contributes most to the immunity-related diseases.[38,39]

The *MHC* genes encode the MHC molecules, which have evolved to maximise the efficacy and flexibility of their functions, in response to a strong evolutionary pressure to eliminate numerous and different types of microorganisms, by binding peptides derived from microbial pathogens and presenting them for recognition by antigen-specific T cells. In all species, there are two types of MHC molecules, known as class I and class II. Both are membrane proteins and contain a peptide-binding cleft at the amino-terminal end. Class I molecules consist of an α chain associated with a β2-microglobulin. The amino-terminal α1 and α2 domains form a peptide-binding groove which is large enough to accommodate peptides of 8–11 residues. The floor of the peptide-binding cleft is the region that binds peptides for display to T cells and the sides and tops are the regions that are in contact with the T cell receptor. Class II molecules consist of α and β chains. The amino-terminal α1 and β1 domains contain polymorphic residues and form a binding groove that is large enough to accommodate peptides of 10–30 residues. The β2 domain contains the T cell co-receptor CD4-binding site.[16]

MHC class I and class II molecules overlap in a number of characteristics: both classes have high levels of polymorphisms, a similar three-dimensional structure and a similar function with regard to peptide presentation at the cell surface of $CD8^+$ cytotoxic and $CD4^+$ helper T cells. However, these molecules have distinct tissue distributions. They differ in the types of antigenic peptide they present: (mainly) intracellular for MHC class I molecules and (mainly) extracellular for MHC class II. In addition, they adopt different pathways. MHC class II alleles are strong genetic susceptibility loci for several autoimmune diseases possible owing to the peptides they present.[38,40,41] MHC class I alleles are also associated with some inflammatory diseases (i.e. ankylosing spondylitis, psoriasis and others) sometimes in interaction with MHC class II alleles (i.e. multiple sclerosis).

An understanding of the genetic complexity of the *HLA* region is also helpful to explain the role of these molecules in the immune response. The *HLA* complex is located on chromosome 6p21.31, containing over 200 defined genes, and can be divided into three classes: class I, class II and class III[38,41] (see Figure2).

The class I region contains approximately 20 class I genes coding for the α polypeptide chain of class I molecules, of which three classic genes (*HLA-A, B* and *C*) are most important. The β2-microglobulin of the class I molecule is encoded by genes located on a separate chromosome. Class I genes are expressed ubiquitously by almost all somatic cells with expression levels varying across tissues.[40]

The class II region contains genes that code for both the α and β polypeptide chains of the class II molecules. Similar to class I genes, three polymorphic genes (*HLA-DR, DQ* and *DP*) are functionally most important. Class II molecules are mainly constitutively expressed by professional antigen-presentation cells, such as dendritic cells, macrophages and B lymphocytes,[40] but their expression can also be induced on many other cells by various stimuli. The prominent SE is encoded by genes within this area, the *HLA-DRB1* alleles.

The class III region occupies a transitional area in between the class I and class II regions and is not structurally or functionally related to either. Instead of possessing direct immune functions, it encodes proteins, such as complement components C2, C4 and factor B, with immune response-related functions. A major role of complement components is to interact with antibody–antigen complexes and mediate activation of the complement cascade, eventually lysing cells, bacteria or viruses.

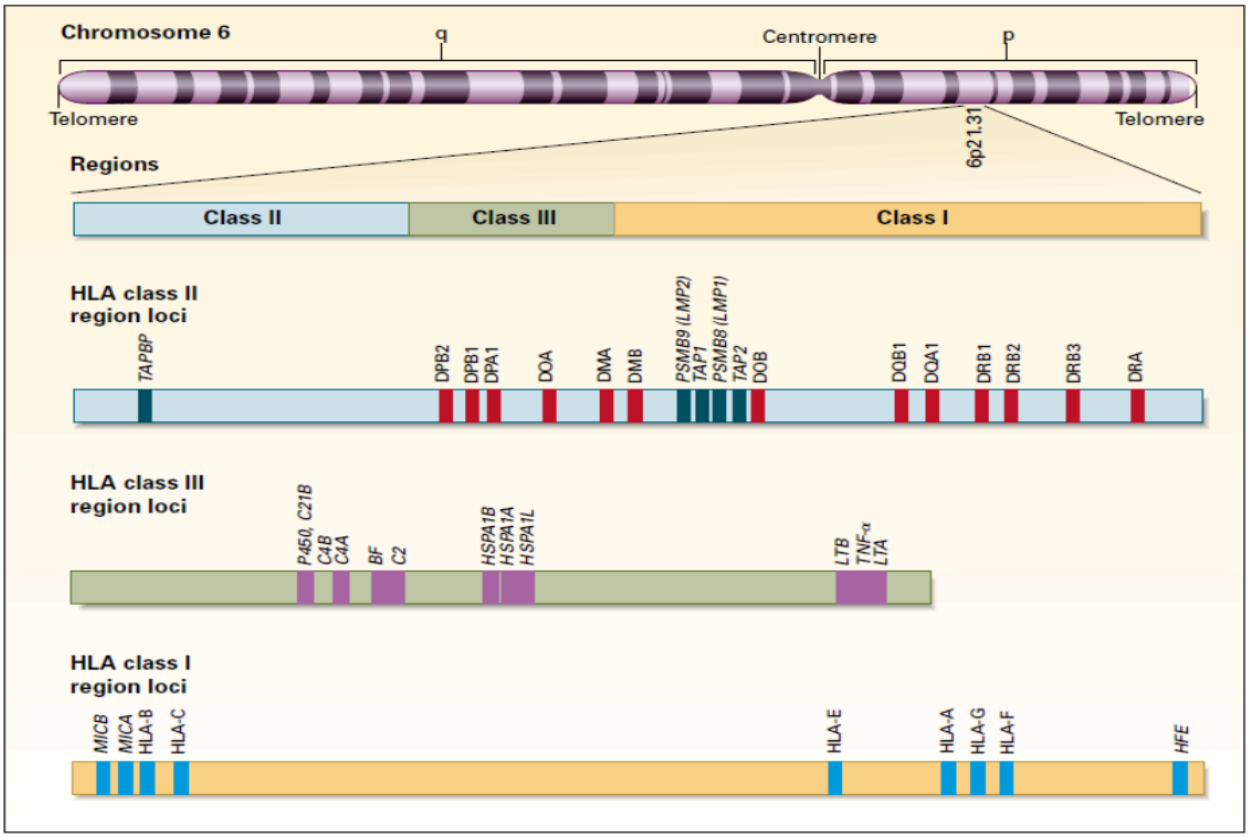


Figure2. Location and organisation of the *HLA* complex on chromosome 6. Adapted from Jan Klein *et al.*[38]

The contribution of *MHC* class II region *HLA-DRB1* gene to RA susceptibility has long been known and is well documented. Peter Stastny first reported in 1976 that HLA-D (and later HLA-DR4) is significantly more common among RA patients than among healthy controls.[42] Subsequently, other HLA-DR serotypes, e.g. the HLA-DR1 in Mediterranean populations or HLA-DR14 in Native Americans, have also been found to be associated with the disease. It has become apparent with the application of modern DNA sequencing techniques since the 1980s that a common feature of the RA-associated HLA-DR molecules is a shared short sequence motif coded by several *HLA-DRB1* alleles. Thus, a "shared epitope hypothesis" was first established by Peter Gregersen *et al.* in 1987.[43] The hypothesis proposes a number of specific *HLA-DRB1* alleles (haplotypes) that encode a conserved sequence motif of five amino acids comprising residues 70–74 in the third hypervariable region of the DRβ1 chain. The three homologous amino acid sequence variants are: 1) QKRAA, the most common motif among Caucasians, coded primarily by the *0401 allele; 2) QRRAA, the second most common motif, coded mainly by *0404, *0101 and *0404; and 3) RRRAA, the least common motif, coded by *1001. The specific SE-coding alleles are shown in Table3.

Table3. Common amino acid sequences in the DRβ 70-74 region.

Amino acid sequence	Shared epitope motif	Coding HLA-DRB1 alleles
QKRAA	+	*0401; *0409; *0413; *0416; *0421; *1419; *1421
DERAA	-	*0402; *0414; *0103; *1102; *1116; *1120; *1121; *1301; *1302; *1304; *1308; *1315; *1317; *1319; *1322; *1416
QRRAE	-	*0403; *0406; *0407; *0417; *0420
QRRAA	+	*0101; *0102; *0105; *0404; *0405; *0408; *0410; *0419; *1402; *1406; *1409; *1413; *1417; *1420
RRRAA	+	*1001
RRRAE	-	*09; *1401; *1404; *1405; *1407; *1408; *1410; *1411; *1414; *1418
DRRAA	-	*0415; *0805; *11011; *11012; *11041; *11042; *1105; *1106; *11081; *11082; *1109; *1110; *1112; *1115; *1118; *1119; *1122; *1201; *12021; *12022; *12031; *12032; *1305; *1306; *1307; *1311; *1312; *1314; *1321; *1601; *1602; *1605
QARAA	-	*15; *1309
QKRGR	-	*03; *0422; *1107

This table is adapted from Joseph Holoshitz.[4]

It has been shown that SE is significantly associated with increased RA risk (especially for ACPA-positive RA) from several independent samples in different worldwide populations.[44-48] In addition to disease susceptibility, SE-coding alleles have also been found to be linked with disease severity and exhibit an allele–dose effect.[49] The mechanism underlying the SE–RA association remains uncertain, but is commonly attributed to the presentation of arthritogenic antigens or T cell repertoire selection.[4]

However, not every RA patient carries SE alleles, and not every SE carrier develops RA, indicating that other factors are important in the disease aetiology. Accumulating evidence has shown the importance of non-SE risk alleles, located from within the same class II region (*DPB1, DOB, DQA, DQB*) extending to the class I region (*HLA-C, HLA-B*), independent of SE in RA aetiology.[50-55] As the technology of imputation has developed, a deeper analysis of amino acids and classical four-digit alleles of *HLA* genes has become possible, allowing researchers to define more clearly the linkage between *HLA* region and RA. Soumya Raychaudhuri *et al.*[56] found the strongest association signal for seropositive RA susceptibility at the HLA-DRβ1 amino acid position 11 (or 13, tightly linked to position 11; both positions are located in the antigen-binding groove and are outside the well-described SE region) but not at the traditional SE positions spanning amino acids 70 to 74. Stepwise conditional analyses identified independent but much weaker association signals at positions 71 and 74. Moreover, positions at the bases of the HLA-B and HLA-DPB1 molecule grooves were also found to confer RA risk. Similarly in East Asians, the same association signal has been observed in ACPA-positive RA among Korean and Chinese populations.[57] Therefore, despite serving as the foundation for RA genetics, SE alone is insufficient to explain the *HLA-DRB1* contribution in RA; neither does it fully explain the SE–smoking interaction often observed in RA.

2.3.3 GWAS in RA

Despite the fact that the most replicable genetic association with RA originates from the most complex region of the human genome, it has been estimated, from the extent of sharing of

identical *HLA* alleles by descent within families, that *HLA* region only accounts for 30% of the total genetic effect.[58] The vast majority of the variance across the genome outside of *HLA* jointly confers the remaining part of the genetic effects. The approaches to measure the genes that underlie common disease and quantitative traits often fall broadly into two categories: candidate-gene studies and GWAS.

Before the advent of GWAS, the candidate-gene approach identified a handful of RA susceptibility loci outside the *HLA* region. These loci included: *PTPN22*,[59] which remains the second strongest RA-associated SNP identified to date, and may act through both T and B cell regulatory activities;[60] protein-arginine deiminase type 4 (*PADI4*);[61] cytotoxic T-lymphocyte protein 4 (*CTLA4*);[62] tumour necrosis factor receptor-associated factor 1 (*TRAF1*);[63] and Fc receptor-like protein 3 (*FCRL3*)[64] (for a summary, see review by Sebastien Viatte *et al*.).[65]

In 1996, the common disease/common variant hypothesis was first proposed, assuming that much of the genetic variation in a common complex disease is due to common variants of relevant small effects.[66,67] It was argued that these common variants would be more easily found by adopting population-based association studies rather than family-based linkage analysis, as the later studies are usually well powered to identify rare variants with large effects. It was also proposed that all common variants in human genes should be recognised, which has been the scientific paradigm for GWAS.

By 2007, GWAS had become feasible due to several crucial advances: 1) the completion of the Human Genome Project providing an accurate blueprint of the human genome sequence; 2) the initial release of the International HapMap project data, depositing millions of genetic markers gathered from four populations (of African, Asian and European ancestry) into the public domain; 3) the availability of information on LD patterns, allowing the design of SNP chips with efficient capture of common variations using only a subset of genome-wide markers (approximately 500000 SNPs); and 4) rapid improvements in SNP genotyping with considerably reduced costs. There are several commercially available GWAS chips, differ in the way in which the SNPs are selected and the total numbers assayed.[68] In the area of immune-mediated diseases, overlapping aetiological factors have long been suggested owing to their shared clinical and immunological features. Therefore, in 2009, investigators of eleven distinct autoimmune and inflammatory diseases (with RA being one of them), designed the Immunochip, an Illumina Infinium SNP microarray interrogate ~190000 SNPs with the major goals as deep replication and fine mapping. The Immunochip has included the top 2000 independent meta-GWAS association signal for each disease, as well as all the SNPs within confirmed GWAS intervals for each disease, without filtering on spacing and LD; and a dense coverage of the *HLA* and killer immunoglobulin-like receptor loci.[68]

From 2007, RA GWAS or Immunochip results have been published almost every year in populations of both European and Asian descent,[69-79] bringing the total number of known RA risk SNPs to 130 (see Table4). Despite this breakthrough, it is generally believed that additional risk alleles for RA remain to be identified.

Table4. List of validated RA susceptibility genes.

Chromosome	SNP	Genes	Populations
1	rs10494360, rs12746613	*FCGR2A*	Korean
1	rs2014863	*PTPRC*	Japanese, Korean, European Caucasians
1	rs2105325	LOC100506023	European Caucasians
1	rs2228145	*IL6R*	European Caucasians
1	rs2240336	*PADI4*	Japanese
1	rs227163	TNFRSF9	Asian
1	rs2476601	*PTPN22*	Korean
1	rs28411352	MTF1-INPP5B	European Caucasians
1	rs2843401, rs3890745	*MMEL1*	European Caucasians
1	rs3753389	*CD244*	Japanese
1	rs3761959	*FCRL3*	European Caucasians
1	rs7537965	GPR137B	Japanese, European Caucasians
1	rs798000, rs11586238	*CD2*	European Caucasians
1	rs883220	*POU3F1*	European Caucasians
2	rs10175798	LBH	European Caucasians
2	rs10209110	*AFF3*	European Caucasians
2	rs11571302, rs3087243	*CTLA4*	Japanese
2	rs11900673	B3GNT2	European Caucasians
2	rs13426947, rs7574865	*STAT4*	European Caucasians
2	rs1980422	*CD28*	Japanese, European Caucasians
2	rs34695944, rs13031237	*REL*	European Caucasians
2	rs6546146, rs934734	*SPRED2*	Japanese, European Caucasians
2	rs6715284	CFLAR-CASP8	European Caucasians
2	rs6732565	ACOXL	European Caucasians
3	rs2062583	ARHGEF3	Japanese, Korean, European Caucasians
3	rs35677470	*DNASE1L3*	European Caucasians
3	rs3806624	EOMES	European Caucasians
3	rs4452313	PLCL2	European Caucasians
3	rs9826828	IL20RB	European Caucasians
4	rs13142500	CLNK	Asian and European Caucasians
4	rs2664035	TEC	European Caucasians
4	rs2867461	ANXA3	Korean
4	rs78560100, rs6822844	*IL2-IL21*	Japanese
4	rs932036, rs874040	*RBPJ*	European Caucasians
5	rs39984	*GIN1*	European Caucasians
5	rs4867947	LCP2	European Caucasians
5	rs657075	CSF2	European Caucasians
5	rs71624119, rs6859212,	*ANKRD55*	European Caucasians
6	rs2234067	ETV7	European Caucasians
6	rs59466457, rs3093023	*CCR6*	European Caucasians
6	rs629326, rs394581	*TAGAP*	Japanese, European Caucasians
6	rs6911690, rs548234	*PRDM1*	Japanese, Korean, European Caucasians
6	rs6920220	*TNFAIP3*	European Caucasians
6	rs9373594	PPIL4	Asian
6	rs9378815	IRF4	Asian and European Caucasians
7	rs3807306, rs10488631	*IRF5*	European Caucasians
7	rs4272	CDK6	European Caucasians
7	rs67250450	JAZF1	European Caucasians
8	rs1516971	PVT1	European Caucasians
8	rs4840565, rs2736340	*BLK*	Japanese
8	rs678347	GRHL2	European Caucasians
8	rs998731	TPD52	European Caucasians
9	rs10739580, rs3761847	*TRAF1*	Korean
9	rs2812378, rs2812378	*CCL21*	European Caucasians
10	rs10795791, rs2104286	*IL2RA*	Japanese
10	rs12413578	10p14	Asian and European Caucasians
10	rs12764378, rs10821944	*ARID5B*	Japanese
10	rs2275806	*GATA3*	Japanese, European Caucasians
10	rs2671692	WDFY4	Asian and European Caucasians
10	rs726288	SFTPD	Asian
10	rs793108	ZNF438	Asian and European Caucasians
10	rs947474, rs4750316	*PRKCQ*	European Caucasians
11	chr11:107967350	ATM	European Caucasians

11	rs3781913	PDE2A-ARAP1	Japanese, European Caucasians
11	rs4409785	CEP57	European Caucasians
11	rs4936059	FLI/ETS1	European Caucasians
11	rs4938573, rs10892279	*DDX6*	European Caucasians
11	rs570676, rs540386	*TRAF6*	European Caucasians
11	rs595158	*CD5*	European Caucasians
11	rs73013527	ETS1	Asian and European Caucasians
11	rs968567	FADS1-FADS2-FADS3	European Caucasians
12	rs10683701, rs1678542	*KIF5A*	European Caucasians
12	rs10774624	SH2B3-PTPN11	European Caucasians
12	rs12831974	TRHDE	Korean
12	rs773125	CDK2	European Caucasians
13	rs9603616	COG6	European Caucasians
14	rs1950897	RAD51B	European Caucasians
14	rs2841277	PLD4	Japanese, Korean, European Caucasians
14	rs3783782	PRKCH	Asian
15	rs8026898	*TLE3*	Japanese
15	rs8043085	*RASGRP1*	Japanese, Korean, European Caucasians
16	rs13330176	*IRF8*	European Caucasians
16	rs4780401	TXNDC11	European Caucasians
17	rs12936409, rs2872507	*IKZF3*	European Caucasians
17	rs1877030	MED1	Asian and European Caucasians
17	rs72634030	C1QBP	Asian and European Caucasians
18	rs2469434	CD226	Asian
18	rs2847297	PTPN2	Japanese, European Caucasians
19	chr19:10771941	ILF3	European Caucasians
19	rs34536443	*TYK2*	European Caucasians
20	rs6032662, rs4810485	*CD40*	Japanese, Korean, European Caucasians
21	rs1893592	UBASH3A	European Caucasians
21	rs2075876	AIRE	Japanese, European Caucasians
21	rs2834512	*RCAN1*	Japanese, European Caucasians
21	rs73194058	IFNGR2	European Caucasians
21	rs9979383	*RUNX1*	Korean, European Caucasians
22	rs11089637	UBE2L3-YDJC	Asian and European Caucasians
22	rs3218251, rs3218253	*IL2RB*	European Caucasians
22	rs4547623	GGA1/LGALS2	Japanese, European Caucasians
22	rs909685	SYNGR1	European Caucasians
X	chrX:78464616	P2RY10	Asian
X	rs13397	*IRAK1*	European Caucasians

2.3.4 Imputation

Despite the tremendous number of genotyped SNPs provided by both the Immunochip and GWAS scan, many SNPs have still not been genotyped. Taking into account that RA is closely linked with the most complex *HLA* region, identifying its precise nature and clearly defining the linkage remain a challenge. The application of imputation has, however, helped to solve this problem to some extent. Imputation is the process of predicting or imputing genotypes that are not directly assayed in a sample of individuals, by comparing the sample of individuals that has been genotyped to a subset of SNPs with a reference panel that has been densely genotyped (or nowadays even sequenced).[80] The theoretical basis of imputation is identical by descent (IBD), indicating that two or more individuals have inherited a segment with the same ancestral origin, so that the segments have similar nucleotide sequences. This is not difficult to understand because, if traced back long enough, all individuals in a finite population are related. Therefore, in samples of unrelated individuals but with the same ethnicity, the haplotypes of the individuals over short stretches of sequence will be related to each other by being IBD.[80] Imputation methods attempt to compare the

underlying haplotypes of the study individuals to the haplotypes in the reference set, identify sharing between the two, and use this sharing to predict missing alleles in the study population. Therefore, it is important when performing imputation to employ large, high-density reference panels that have ancestry close to that of the study population. HapMap 2 haplotypes and large reference panels from the other consortia have been widely used. The algorithms of imputation differ in the way in which specific models are used, but all essentially comprise a phasing procedure, after which the haplotypes are compared to and modelled as a mosaic of the dense haplotypes in the reference panel (see Figure3). Missing genotypes are then imputed through matching. Imputation can be implemented within a focused chromosome region, or over the genome. It helps to perform fine-mapping, to harmonise data for meta-analysis and to correct genotype errors, which all boost the power of the study.

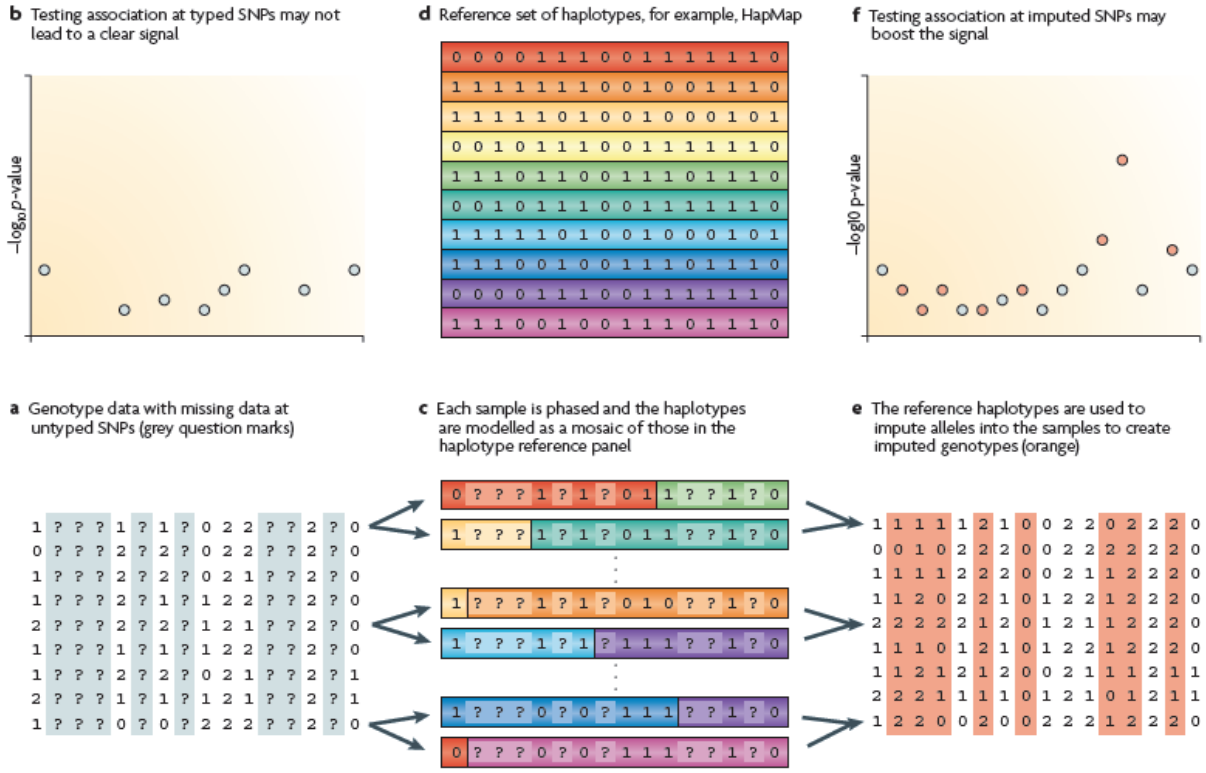


Figure3. The method of genotype imputation. Adapted from Jonathan Marchini *et al*.[80]

2.3.5 The Concept of Heritability

Before considering heritability estimates in RA, it is useful to understand the concept of heritability. Here height will be taken as an example to illustrate the relevant factors. Assuming in a certain population that body height has been measured in everyone, it is reasonable to expect that the measurements would be diverse: some individuals would be taller while others would be shorter. Thus the body heights of this particular population could be described by a mean (μ) and a variance (σ^2) indicating the spread of the set of numbers (i.e. variance equal to 0 indicates identical values for the entire sample). A question that might be of interest is whether the observed variation in height is due to environmental or genetic factors: the nature–nurture debate (which is a simplified way of ignoring the

genotype–environment correlation or interaction). According to this idea, the variance could thus be divided into two parts, attributed to genotype (σ^2_G) and environment (σ^2_E); the genetic variance could be further subdivided into the variance of additive genetic effects (σ^2_A, i.e. each allele adds some effect to the phenotype), dominant genetic effects (σ^2_D; of note, all effects that are non-additive are considered dominant in this case) and gene–gene interactions (σ^2_I). Therefore the broad-sense heritability (H^2) is defined as σ^2_G / σ^2, and the narrow-sense heritability (h^2) as σ^2_A / σ^2. In practice, narrow-sense heritability is usually referred to, because it is the chief cause of resemblance between relatives and the main determinant of selection response (see Figure4). To summarise, heritability is the extent to which differences in the phenotypes of a trait can be accounted for or predicted by differences in genes;[81] that is, the additive genetic variation. Gene–gene or gene–environment interactions are not considered in this very simplified definition.

Several misconceptions regarding heritability should be considered. Firstly, heritability cannot be applied to an individual person because there will be no variance, therefore it does not indicate the proportion attributable to genetic factors for the traits of an individual. Secondly, heritability does not reflect the extent of a phenotype that is passed on to the next generation; this is only determined by genes. A high heritability implies that most of the variation observed is caused by variations in genotype, but it does not mean that the phenotype is determined once the genotype is known, nor does it determine how modifiable a trait is to environmental influences. Furthermore, it is well known that the heritability of human height is 80–90%. This does not mean that 1) height in all individuals is determined by 80–90% by genetics and 10–20% by the environment, 2) that environmental intervention should not be employed to increase body height as "it does not contribute much" or 3) that by removing all the "short genes", 80–90% of the "short stature" of the population could be improved. Finally, heritability varies across populations, thus within-group heritability in theory cannot be used to explain between-group differences but, in reality, this is usually done because the heritability for some diseases remains stable across the population. Heritability can also differ between sexes and with age. See review by Peter Visscher *et al*. for a detailed description of these concepts.[81]

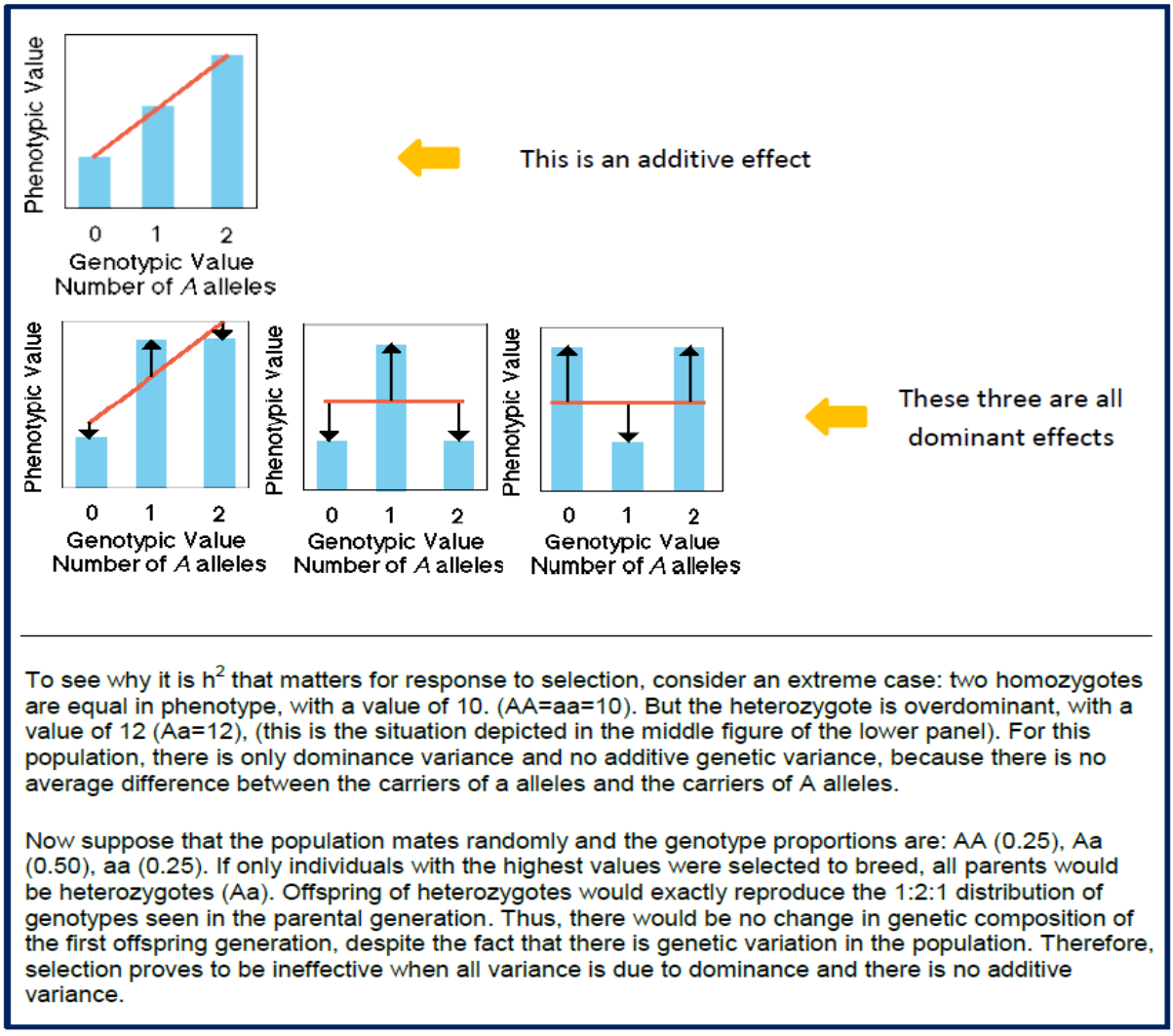


Figure4. Additive variance versus dominant variance: why narrow-sense heritability is most commonly referred to?

Disease status is usually defined in a binary fashion, rather than as a continuous variable. When the concept of genetic heritability, which is usually applied to depict variation in quantitative traits (such as body height and body weight), is extended to include categorical traits (RA versus non-RA), a genetic liability model is generated. It is assumed that there is an underlying, unmeasured yet normally distributed continuous liability scale for a phenotype, with the presence or absence of the phenotype determined by a threshold. Weight is a good example of such a scenario: clinically, individuals could be divided into either obese or non-obese, however the binary trait is based on the continuous scale of body mass index, which serves as the underlying unobserved liability scale. Such liability scales exist in all other diseases. Heritability in this case is estimated from the pattern of associations in categorical traits measured in first degree relatives (FDRs).

Traditionally, heritability is estimated from the degree of resemblance between relatives, e.g. parents–offspring, siblings and twins. Studies usually adopt simple and balanced designs based on families, such as regression functions of offspring on parental phenotypes, correlation of full or half siblings and difference in the correlation between monozygotic and

dizygotic twin pairs.[81] These methods are straightforward: the phenotype of the two relatives will always be positively correlated as they share genetic causes of the phenotype to different extents, depending on their kinships. In addition, the more a trait is aggregated within families, the higher the heritability owned by that trait. In the era of genomics with the development of high-density genotyping platforms, estimation of heritability in unpedigreed populations for complex or medically important traits becomes possible. Genetic markers can be used to calculate the relatedness between pairs of individuals and help to estimate heritability in a novel way.[81]

2.3.6 Heritability in RA

What is the level of heritability in RA? The results of several early twin studies have indicated that RA heritability is 40–60%.[82-84] In a recent study of 12590 twins, ACPA-positive RA heritability was calculated to be 41%.[85] Studies using genome-wide markers have estimated the heritability of ACPA-positive RA to be 40–55%.[86,87] In a large population-based study of FDRs, it was estimated that the heritability of ACPA-positive and ACPA-negative RA was ~50% and ~20%, respectively.[6] There is some variation regarding the purported strength of familial aggregation of RA. The results of hospital-based surveys, with limited sample sizes and lacking controls groups, have suggested that 7–22% of RA patients have one or more FDRs affected by the disease.[88-92] Studies with representative control groups have reported relative risks (RRs) ranging from 1.7 to 4.6.[93-97] A recent register-based study in the Swedish total population found familial odds ratios (ORs) for RA of about 3 in FDRs and 2 in second-degree relatives.[6]

The extent to which identified genetic loci may explain RA heritability has been explored in GWAS, and several authors have concluded that identified genetic loci account for about half of RA genetic liability,[77,98,99] or 50–65% and 30–50% of the genetic liabilities of ACPA-positive and ACPA-negative RA, respectively.[100] However, identified non-HLA alleles only explain 5–14% of seropositive RA heritability.[75,77,79,101] It was recently calculated that identified variants in the *HLA* region would explain 12.7% of ACPA-positive RA liability.[56] These estimates seem to indicate roughly equal importance of genetic and environmental factors for the development of seropositive RA. It appears that for the development of seronegative RA, environmental factors are of relatively greater importance than genetics, but it should be noted that seronegative RA also harbours greater heterogeneity. Moreover, the estimates further imply that identified genetic markers should explain a substantial proportion of the familial aggregation of RA.

2.4 ENVIRONMENTAL FACTORIS IN RHEUMATOID ARTHRITIS

Emerging data suggest that the systemic inflammation and autoimmunity in RA begin long before the onset of detectable joint inflammation. Environmental factors might play an important role as the first triggers of RA at sites distant from the joints (i.e. the lung, oral cavity and gut) where it is believed that inflammation originates.[102] The established environmental risk factors for RA will be introduced briefly in this section.

2.4.1 Smoking

So far, the most clearly demonstrated environmental risk factor of major importance for RA is smoking, and this has been consistently replicated in a number of independent epidemiological investigations in a variety of populations of different ethnicities.[7-11,103,104] Two recently published meta-analyses have demonstrated that ever, current and past smokers carry an OR of 1.9 (95% confidence interval (CI): 1.6–2.3), 1.9 (95% CI: 1.5–2.3) and 1.8 (95% CI: 1.3–2.3) for the risk of developing RA, respectively,[105] and that the increased risk effect is dose dependent, where RA risk is elevated by smoking by 26% for 1–10 pack-years, 94% for 20–30 pack-years and 107% for >40 pack-years.[106] The effect of smoking has been found to be restricted to ACPA-positive RA.[10,103] Less is known about the impact of smoking cessation on the future risk of RA, with a few studies finding a reduction, but not elimination, in risk after 10–20 years of smoking cessation.[7-9]

Cigarette smoke is a complex mixture consisting of various chemical compounds, e.g. nicotine, char and adjuvants, each of which has a different function on the immune system.[15] Some constituents such as char may act as adjuvants and enhance both innate and adaptive immunity by activation of antigen-presenting cells[107] whereas nicotine may suppress inflammation through relatively well-investigated mechanisms.[108] Therefore it is of interest to investigate which substances in inhaled smoke increase RA risk: the particles or the nicotine.

2.4.2 Other Airway Exposures

In addition to smoking, another well-characterised inhalation exposure for RA is silica, which has long been known to be an occupational hazard of working in certain industries. A moderate risk effect of silica dust has been observed for ACPA-positive (OR: 1.6) and RF-positive RA (OR: 1.9).[109-111] An additive interaction between silica and smoking has also been reported, with an OR of 7.36 among silica-exposed current smokers and an attributable proportion due to interaction of 60%, in the aetiology of ACPA-positive RA.[110] Other respiratory factors, such as mineral oil, air pollution and organic solvents, were also reported to increase RA risk with marginal significance.[112-114]

2.4.3 Alcohol

A significant inverse association between alcohol consumption and RA risk for both men and women has been observed in a number of case–control and cohort studies.[115-119] Compared with individuals who do not drink any alcohol, a 30–40% dose-dependent decreased risk has been identified among moderate drinkers. The adverse impact has also been demonstrated in terms of disease severity.[117] A possible explanation of the underlying mechanism could be that alcohol exerts effects on both the hormonal and the immunological systems to downregulate immune response, suppress the synthesis of proinflammatory cytokines/chemokines and elevate serum oestradiol concentrations.[115] A complementary molecular explanation has been suggested from studies on experimental arthritis in rodents where alcohol reduces inflammation and arthritis incidence and severity through inhibition of the NFKB pathways.[120]

2.4.4 Other Lifestyle-related Factors

Several lifestyle-related factors associated with moderate risk of RA have been well established, in addition to cigarette smoking and alcohol consumption. Being overweight/obese is associated with an increased RA risk, a worse long-term disease outcome, and a poor response to certain treatments.[121-123] This association might be due to the immunomodulating and proinflammatory properties of adipose tissue. Dietary habits such as low or no fatty fish consumption are also associated with increased RA risk.[124-126] Furthermore, low socioeconomic status and high levels of personal and neighbourhood deprivation have been linked to increased RA risk.[127,128] Finally, hormonal and reproduction-related factors such as parity, breastfeeding and oral contraceptive use are important RA-related factors among women, although inconclusive results have been reported.

2.5 GENE–ENVIRONMENT INTERACTIONS IN RA

2.5.1 Concept of Interaction

In many settings, the effect of one exposure may depend in some way on the presence or absence of another exposure; in other words, there is an interaction between the two exposures. The term "interaction" has different meanings which easily generates confusion. The interpretation thus largely depends on subjective opinions. Terms such as "effect modification", "effect measure modification", "synergism" and "antagonism" have all been used to describe interaction.[129] Interaction can occur between genetic and environmental exposures, as well as between two (or more) environmental exposures, or (two or more) genetic exposures.

From a statistical point of view, interaction refers to the necessity of adding a product term in a statistical model in order to "fit the data well".[130] The product terms can be included in different types of models. The two most commonly used statistical models in epidemiology are logistic regression and Cox regression, which are inherently multiplicative (they are implicitly exponential). Thus, the presence of an interaction term in such a model implies departure from multiplicativity; whereas the inclusion of an interaction term in a linear regression model implies departure from additivity. Therefore, whether statistical interaction reflects departure from multiplicativity or additivity depends largely on the selected models. Because the vast majority of epidemiological analyses are based on multiplicative models, most results utilise and reflect the multiplicative scale. Moreover, statistical interaction may be applied solely on the basis of concern for statistical convenience, e.g. to fit the data best, without any consideration for biological mechanisms. The strengths of statistical interaction are the simple description of observed phenomena and often with prediction as a related goal, which do not need to be linked with biological inference (e.g. multivariate risk predictive functions).[131]

In the field of epidemiology, it has long-been debated whether the scale of interaction should be determined by the best-fit statistical model considering only the statistical convenience

and simplicity, or whether it should be assessed in a model that corresponds well to the biological process to also incorporate biological accuracy.

Biological interaction refers to the mutual or co-dependent operation of two or more causes to produce the disease, which is the case for most diseases. It has been argued that the additive scale is more appropriate to evaluate biological interaction, and that it fits with the sufficient-component concept of causality, as suggested by Rothman.[132] Sufficient causes could be interpreted as minimal sets of actions, events or states of nature that together initiate a process that inevitably results in the outcome.[132] For a particular outcome there would be many sufficient causes, each involving various component causes. If two independent causes are both the components of the same sufficient cause, then they participate together and a synergistic effect is presented.[133] Additive interaction could thus be defined as the synergistic effect of two contributory causes of a disease exceeding the sum of their independent effects.[131] Often the additive interaction is of greatest public health importance in that identification of all the components of a given sufficient cause is unnecessary for prevention; blocking the causal role of one component renders the joint action of the other component(s) insufficient, and thus prevents the effect. Three measures have been proposed to assess additive interactions: the relative excess risk due to interaction (RERI), the attributable proportion due to interaction (AP) and the synergy index (S).[134]

Assuming that two exposures contribute to the disease, and RR_{00} is the relative risk of individuals with none of these exposures, RR_{01} and RR_{10} are the risks of individuals with either one exposure and RR_{11} is the risk of individuals with both exposures, the three measures can be calculated as follows:

$$RERI = RR11 - RR10 - RR01 + 1;$$

$$AP = \frac{RERI}{\mathrm{RR11}};$$

$$S = \frac{\mathrm{RR11}-1}{(\mathrm{RR10}-1)+(\mathrm{RR01}-1)}.$$

2.5.2 SE–Smoking Interaction in RA

A striking gene–environment interaction between smoking and SE was first reported by researchers from Sweden (our group). This work demonstrated a strong interaction effect only associated with seropositive RA.[14] Furthermore, the additive interaction suggested a biological pathway to disease onset. These findings have been replicated and expanded in subsequent studies.[135-138] In an attempt to identify specific mechanisms by which SE and smoking trigger RA, bronchoalveolar lavage (BAL) fluid has been examined from healthy non-smokers, health smokers and smokers with inflammatory lung diseases. It was found that both the expression of PAD2 enzyme in human lungs and the proportion of citrulline-positive BAL cells were elevated by smoking, whereas citrullinated cells were not present (to the same degree as smokers) in non-smokers.[13,139] These findings led to the hypothesis that smoking induces citrullination of proteins in the lungs, as well as the prevalence of inducible

bronchus associated lymphoid tissue. Both mechanisms can result in a production of RF and ACPA. The antibodies, together with a correct genetic background (SE and *PTPN22*) and another trigger in the synovium, lead to the development of RA. In addition to *SE*, smoking interacts with several well-characterised loci, including *TNFAIP3, CTLA4, STAT4, PTPN22, TRAF1/C5, PADI4, GSTT1, GSTM1, HMOX1* and *EPXH1*.[140-142] However, these previous studies of the interplay between smoking and genetics used the candidate gene approach instead of examining the whole genome, and the results were mostly restricted to one population.

3 AIM

3.1 OVERALL AIM

To study the gene–smoking interaction in the aetiology of RA.

3.2 SPECIFIC AIM

Study I: To investigate the cigarette smoking–SNP interaction in RA by using markers from both GWAS and Immunochip materials, and to identify novel variations within and outside the HLA region.

Study II: To explore whether exposure to nicotine is associated with RA risk, using snuff consumption as an indicator of nicotine exposure.

Study III: To clearly define the interaction between cigarette smoking and HLA amino acid positions in seropositive RA, with the new HLA-DRβ1 amino acid model.

Study IV: To test how much of the RA familial risk can be explained by currently established genetic and environmental risk factors.

Table5. Summary of the material used in the four studies.

Study	Datasets				Nationwide registers		Genetic materials		Imputation	Environmental factors	Outcomes
	EIRA Case-control design	Umeå Nested case-control design	NHS Nested case-control design	Korean Case-control design	The Multi-generation Register	The Swedish Patient Register	Immunochip	GWAS			
Study I	√ (Discovery dataset, both EIRA I and II)	√ (Replication dataset)					√ (Whole Immunochip material used)	√ (Whole GWAS material used)	√ (HLA alleles were imputed on Umeå subjects)	Cigarette smoking (Status)	RA, ACPA-positive and ACPA-negative RA
Study II	√ (EIRA I)									Snuff use	RA, ACPA-positive and ACPA-negative RA
Study III	√ (Both EIRA I and II)		√ (Both NHS I and II)	√			√ (Some markers on chromosome6 from the Immunochip were used to impute amino acids and HLA alleles)		√ (HLA alleles and amino acids were imputed on both EIRA and NHS subjects)	Cigarette smoking (status and intensity)	Seropositive RA
Study IV	√ (Both EIRA I and II)				√ (Identify EIRA subject's FDRs)	√ (Identify the RA status among FDRs)	√ (Currently identified risk SNPs were extracted from the Immunochip)			currently known environmental factors	RA, ACPA-positive and ACPA-negative RA

4 MATERIALS AND METHODS

Three of the four studies included in the current thesis have incorporated more than one independent dataset; an overview has been presented in Table5. Briefly, four datasets and two nationwide registers were involved. Study I comprised a discovery phase (EIRA) and a replication phase (Umeå) implemented in two separate datasets; Study II was based on EIRA alone; Study III was carried out in parallel in three populations (EIRA, NHS and a Korean cohort); and Study IV was largely based on EIRA with linkage to two nationwide registers. All the studies were of case–control design with exposure and outcome variables dichotomised. Therefore, logistic regression was primarily employed in terms of statistical models. The two main analyses, an association and an interaction analysis (additive and multiplicative, respectively), investigated RA overall as well as RA subsets according to serotypes. All the studies were approved by the relevant regional ethics committees and all subjects provided informed consent to participate.

4.1 MATERIALS

4.1.1 Study Design and Populations

The EIRA Study

EIRA, a population-based case–control study initiated in 1996, includes incident RA cases and controls aged 18–70 years recruited from defined (southern/central) regions of Sweden. EIRA is actively enrolling new subjects. The data based on the first version of the EIRA questionnaire during the recruitment period between 1996 and 2006 is defined as EIRA I. From 2006, a second recruitment period with a modified questionnaire was initiated and EIRA II denotes the data from this second phase. Cases were patients who received a diagnosis of RA according to the 1987 ACR criteria by rheumatologists in the study area. The aim of EIRA is to identify incident cases in the study database as soon as possible after the onset of disease. Therefore the median symptom duration for more than 90% of EIRA cases is 10 months. Controls were randomly selected, matched for age, gender and residential area with the cases, from a continuously updated national register. One control (two controls since 2006) was selected per case. Details of the study design have been described elsewhere.[11] Subjects filled out a self-administered questionnaire at baseline, and blood samples were collected for further genetic or serological tests.

The Umeå Study

The Umeå study population was based on a nested case–control study established through two population-based cohorts (the Northern Sweden Health and Disease Study cohort, and the Maternity cohort of Northern Sweden) in the four northern-most counties of Sweden. All eligible patients were diagnosed with early RA according to the 1987 ACR criteria, and were consecutively included by rheumatologists at the Department of Rheumatology, University Hospital Umeå. Controls were randomly selected from the Medical Biobank of northern Sweden, matched for age and gender with the cases. Four controls were selected per case.

Details of the study database have been reported elsewhere.[143] Subjects filled out a self-administered questionnaire at baseline, and provided blood samples for further genetic or serological measurements.

The NHS

The NHS was established in 1976 in the USA. Registered nurses, aged 30 to 55 at the time, who lived in the 11 most populous states and whose nursing boards agreed to contribute to the study, were enrolled in the cohort. These nurses were required to respond to the baseline questionnaire, and were selected for prospective follow-up. Every 2 years they receive a follow-up questionnaire with questions about diseases and health-related topics (e.g. smoking, hormone use and menopausal status). Blood samples were collected to identify potential biomarkers. The NHS subjects included in the current thesis were from a nested case–control study based on the prospective NHS. Women who reported RA were screened for RA symptoms; chart review confirmed RA according to the 1987 ACR criteria. Healthy control subjects were selected and matched with the cases at the index date of diagnosis by age, menopausal status and post-menopausal hormone use.

The Korean Study

The Korean cases were recruited from Hanyang University Hospital for Rheumatic Diseases in Seoul, Korea, and were all ACPA positive. The controls were healthy volunteers recruited partly from the hospital (healthy hospital staff), and partly from recruiting campaigns. Smoking information was retrospectively recorded as never, past or current smoking and, for smokers, pack-years of smoking at RA onset.

The Swedish Nationwide Registers

The index persons in the Swedish Multi-Generation Register are Swedish residents born in 1932 or later, and registered as alive in 1961. The register identifies the parents of the index persons.[144] With information available on more than 9 million individuals, the coverage is good from the 1940s and almost complete for those born in or after 1968. FDRs of EIRA subjects were ascertained through this register.

The Swedish Patient Register contains information about inpatient treatment since 1964 (complete nationwide coverage from 1987), and outpatient visit diagnoses from non-primary care since 2001.[145] RA was defined as having more than two diagnoses of RA, with at least one assigned by specialists in rheumatology or internal medicine. Validation studies have demonstrated that approximately 90% of such individuals, whether admitted to hospital or diagnosed during an outpatient visit, fulfill the 1987 ACR criteria.[146] RA occurrence among the FDRs of EIRA subjects was assessed through this register.

4.1.2 Genetic and Biological Measurements

Enzyme-linked Immunosorbent Assay

Serum ACPA status was measured using an enzyme-linked immunosorbent assay (ImmunoscanCCPlus, Euro-Diagnostica). Cut-off for positivity was 25 U/ml. RF status was assessed based on the patients' history of seropositive or seronegative RA, according to the 10^{th} revision of the International Classification of Disease codes. Both ACPA and RF data were available for EIRA and Umeå study participants, whereas only RF was available in NHS.

SE Genotyping

The genotyping procedures for *HLA-DRB1* alleles were performed in blood DNA samples by sequence-specific primer-polymerase chain reaction. For the *HLA-DRB1* low-resolution kit, an interpretation table was used to determine the specific genotype according to the manufacturer's recommendations. *HLA-DRB1*01* (except *DRB1*0103*), **04* and **10* were classified as SE alleles.

Immunochip

Immumochip was used as a source of genetic markers, as well as the material for imputation. The EIRA Immunochip scan included 195586 genetic markers from 5043 samples. Data were filtered on the basis of both SNPs and individuals (SNP genotype call rates >95% completeness in both cases and controls; minor allele frequency >0.01 in both cases and controls; and Hardy-Weinberg equilibrium p-value $>1\times10^{-5}$ in controls; in addition, subjects with more than 5% missing genotypes, evidence of relatedness and non-European ancestry were excluded).

Population stratification was controlled using the principal component approach (PCA) implemented in software EIGENSTRAT. The high-density genotype data provided by the international HapMap project release III for four reference populations, CEPH trios from Utah with Northern European ancestry, Yoruba trios from Ibadan Nigeria, unrelated Japanese individuals from Tokyo and unrelated Han Chinese individuals from Beijing, were downloaded and used as the reference sample. PCA analysis was performed for the EIRA sample combined with the reference sample at the same available SNPs (n=44863). Outliers were identified and deleted. This trimming step was iteratively executed, removing 17 outliers in five iterations. After quality control (QC), a total of 133648 SNPs and 4337 participants (1590 ACPA-positive RA patients, 891 ACPA-negative RA patients and 1856 control subjects) were included.

After similar QC, Immunochip data were available for 1859 individuals (614 ACPA-positive and 271 ACPA-negative RA patients, and 974 control subjects) in the Umeå study, and 598 individuals (235 seropositive RA patients and 363 control subjects) in the NHS study.

GWAS Scan

The Immunochip primarily targets the known immunity-related genes, and is therefore not truly "genome wide". Thus EIRA GWAS data, obtained using the Illumina 300K chip and

available for 1147 ACPA-positive and 774 ACPA-negative RA cases and 1079 controls for 301171 markers after QC, were included as another source of genetic markers.

Imputation

The procedure for HLA imputation was performed using HLA2SNP software according to the manufacturer's instruction manual. Immunochip data were used, together with a publicly released European reference panel generated by the Type1 Diabetes Genetics Consortium (T1DGC). This reference dataset contains SNPs selected to cover the entire *MHC* region and *HLA* alleles at four-digit resolution in 2767 unrelated individuals of European descent. Missing genotypes in the Immunochip sample, as well as the subsequent amino acids, were imputed using those matching haplotypes in the reference panel. For the EIRA and NHS datasets, the classical alleles and amino acids of HLA-DRB1 were imputed. For the Umeå dataset, only the SE alleles were imputed. The concordance rate between imputed SE and genotyped SE at two-digit resolution were 97.4%, 97.3% and 93.8% for the Umeå, EIRA and the NHS cohorts, respectively.

4.1.3 Environmental Factors

Exposure to cigarette smoking was based on self-reported questionnaires. Ever smokers were defined as individuals who reported that they smoked or had previously smoked cigarettes (current smokers and former smokers); never smokers were defined as those who reported that they had never smoked cigarettes. Subjects who smoked pipes or cigars were excluded, thus the ever smoker group was restricted to cigarette smokers. Only exposure data up to the index year (the year in which first RA symptoms occurred in cases, and the same year for the corresponding controls) was used. Pack-years of smoking were calculated with 1 pack-year equivalent to smoking 20 cigarettes per day for 1 year, and were categorised as below/equal to versus above 10 pack-years.

Exposure to snuff use was based on information from a self-reported questionnaire. The relevant section contains two questions regarding snuff use: 1) Are you currently using snuff? Yes/No; and 2) If not, have you previously used snuff? Yes/No. Ever snuff users were defined as individuals who reported both current and former snuff use; never snuff users reported that they had never used snuff.

Other lifestyle-related factors were categorised adopting the same categorisation as in our previous publications. Briefly, ever drinking was defined by questions about present alcohol consumption as well as previous habitual consumption, including both current and former drinkers. Total alcohol consumption was calculated based on drinks per week (with one drink equal to 16 g alcohol). Ever parious was only relevant among women and was defined as having ever given birth. Individuals who were construction workers were included as silica exposed; and whose work involved either rock drilling or stone crushing in particular were considered to have been exposed to high levels of silica. Body mass index was calculated based on self-reported height and weight. Fatty fish consumption was categorised into frequent (at least 1–2 times/week), less frequent (at least 1–2 times/month but less than 1-2

times/week) and never/seldom consumption. Level of education was determined through linkage to the Swedish Register of Education, and was classified as ≤9, 10–12 and 12+ years.

4.2 STATISTICAL ANALYSIS

4.2.1 Study I

Exposures: Cigarette smoking status (ever/never), genetic markers from the Immunochip (risk allele/referent allele) and genetic markers from the GWAS (risk allele/referent allele).

Outcomes: ACPA-positive and ACPA-negative RA.

Statistical analysis: Additive and multiplicative interactions between smoking and SNPs in the risk of developing ACPA-positive and ACPA-negative RA.

We first calculated the AP together with its 95% CIs to evaluate the additive interaction between smoking status (never/ever) and the Immunochip markers, adjusted for the matching factors (age, gender and residential area). A Bonferroni corrected p-value for the AP of less than 0.05 (corresponding to a genome-wide p-value threshold of 3.74×10^{-7}) was set as the threshold for significance. Three genetic models (dominant, recessive and co-dominant models) were applied. To maintain the stability of our results, smoking–SNP pairs with a minimum cell frequency of less than 5 were considered unreliable and were not included in the final results regardless of p-values. We then performed the same analysis among those significant smoking–SNP pairs with adjustment of copies of SE, to identify markers independent of SE. Each step was replicated in the same way using the data from the Umeå study. We also evaluated the multiplicative interaction by adding a product term in the logistic model. To further extend our findings, we included EIRA GWAS data and performed smoking–GWAS interaction analysis in order to identify any possible signals that might have been neglected due to the particular design of the Immunochip. All analyses were performed using SAS version 9.3 of the GEIRA program.[147]

4.2.2 Study II

Exposure: Snuff use (ever, current and former users versus never users).

Outcome: Overall, ACPA-positive and ACPA-negative RA.

Statistical analysis: Association between snuff use and RA risk.

We calculated the ORs and 95% CIs for the risk of RA overall, as well as for ACPA-positive and ACPA-negative RA, associated with snuff use, through unconditional logistic regression models. Exposed groups (ever, current and former snuff users) were compared with reference group (never users). First we performed the association analysis based on all cases and controls, with adjustment for the matching variables (age, gender and residential area), for cigarette smoking and for alcohol consumption. We further performed the same analysis with matching using conditional logistic regression models. Additional adjustments for education, silica exposure, body mass index, the primary genetic risk factor SE and *PTPN22* gene did

not alter the ORs substantially and were therefore not retained in the final models. Because cigarette smoking is the major RA environmental risk factor, and might be an important confounder, we also performed an analysis stratified by smoking status (never smokers and ever smokers, respectively). We further carried out the above-mentioned analyses among men and women separately. SAS version 9.3 was used for all procedures.

4.2.3 Study III

Exposures: Amino acid positions 11, 13, 71, 74 at HLA-DRβ1, position 9 at HLA-B and position 9 at HLA-DPβ1 (carriers of certain residue versus non-carriers). Genetic risk score (GRS, dichotomised at the median of the controls) calculated for each amino acid positions as well as for haplotypes (based on the four amino acid positions at HLA-DRB1). Cigarette smoking status (never versus ever smokers) and intensity (≤10 versus >10 pack-years of smoking).

Outcome: Seropositive RA (ACPA or RF positive).

Statistical analysis: Association analysis between amino acids and RA, and between smoking and RA; interaction analysis between amino acids GRS/haplotype GRS and smoking in RA risk.

We first calculated the GRS for individual haplotypes by weighting each haplotype based on its reported RA risk effect size for Europeans according to data from Raychaudhuri *et al.*,[56] using the following equation:

$$GRS_i = \sum_{h=1}^{n} (\ln OR_h) \cdot haplotype_{hi}$$

where GRS_i is a haplotype GRS of individual i, OR_h is the reported OR of haplotype h and $haplotype_{hi}$ is the number of haplotype h in individual i.

We also calculated the GRS for each of the amino acid positions using the following equation:

$$GRS_{ki} = \sum_{a=1}^{n} (\ln OR_{ak}) \cdot Dosage_{aki}$$

where GRS_{ki} is a GRS of amino acid position k in individual i, OR_{ak} is the reported OR of residue a at amino acid position k (6, 6, 4, 5, 3 and 3 residues at the positions HLA-DRB1 11, 13, 71, 74, HLA-B 9 and HLA-DPB1 9, respectively) and $Dosage_{aki}$ is the imputed dosage of residue a at amino acid position k in individual i. The GRS was dichotomised at the median of the controls for the subsequent interaction analysis.

We first calculated the OR and 95% CIs of heavy smoking on RA risk by logistic regression with adjustment for age, gender and residential area. Associations between the amino acid positions and RA risk were evaluated by omnibus test with p-values calculated by log-

likelihood ratio tests comparing the fit between the null model and full model. We subsequently assessed the additive interaction by calculating AP and its 95% CIs, with two principal components used as covariates. For each of the amino acid positions' GRS, we additionally conditioned on GRSs for other amino acid positions to remove the correlation effect generated by LD among the amino acid positions. The amino acid positions 11 and 13 are tightly linked, therefore were not adjusted with regard to each other in the AP model. The multiplicative interaction was assessed in a logistic regression by a product term. All analyses were performed using SAS version 9.3.

4.2.4 Study IV

Exposure: FDRs with versus FDRs without a RA diagnosis. The RA status among FDRs was further categorised as RF-positive and RF-negative RA.

Outcome: Overall RA, seropositive (ACPA or RF) and seronegative (ACPA or RF) RA.

Statistical analysis: Association between family history of RA and RA risk, with adjustment for known genetic and environmental RA risk factors.

The strength of familial aggregation was evaluated as the OR for the association between the "exposure" (RA among FDRs) and the outcome (RA among EIRA subjects), estimated using unconditional logistic regression. ORs were first estimated in a crude model with adjustment for the design variables (age, gender and residential area) alone, and then in separate adjusted models for each (or groups of) risk factor(s). In addition to models for specific covariates, we fitted models simultaneously adjusting for individual SNPs, the GRS (calculated based on SNPs) and SE, as well as a model adjusting for all environmental factors, and a full model adjusting for all covariates. We performed stratified analyses by ACPA status (and RF status) in index cases and by RF status in FDRs. All analyses were performed using SAS version 9.3.

To accommodate the missingness and the potential selective participation observed among controls, multiple imputation (MI) using fully conditional specification models was utilised. The imputation model included all the variables in the dataset, with linear and squared terms for continuous covariates, and predicted missing values for any variable using existing values from other variables. To confirm that the imputation did not bias the results, we also performed a complete-case analysis, in which the unadjusted and adjusted model for each covariate was fitted only among individuals who had complete information on that particular (or group of) covariate(s).

5 RESULTS

5.1 STUDY I

After QC, we identified a set of 133648 SNPs with an average call rate of 99.7% in 2481 incident RA cases (1590 ACPA-positive and 891 ACPA-negative cases) and 1856 controls from the EIRA study; in the Umeå dataset, 885 RA cases (614 ACPA-positive and 271 ACPA-negative cases) and 974 controls were available. We did not observe any heterogeneity between the EIRA and Umeå data in terms of population stratification (see Figure5).

In the EIRA study using the Immunochip data, a total of 102 non-duplicated SNPs were found to significantly interact with smoking in ACPA-positive RA; all were located in the *HLA* region (one from *HLA* class I and the rest from *HLA* class II). Among these 102 SNPs, three were located in the gene-coding region of chromosome 6 (rs3749966 in *C6orf10*, and rs8084 and rs7192 in *HLA-DRA*) and four were located in the 3' untranslated region (rs7195, rs1051336 and rs1041885 in *HLA-DRA*, and rs10484565 in *TAP2*), while the rest were from intergenic/non-coding regions. The lack of interaction signals outside the *HLA* region might be due to the selective coverage of Immunochip genetic markers. We therefore performed GWAS–smoking interaction analysis but no additional signals outside chromosome 6 were identified. By contrast, no SNP was found to statistically significantly interact with smoking in ACPA-negative RA in any genetic models. Subsequently, we restricted our analyses solely to ACPA-positive RA cases and controls.

We then analysed all the 102 SNPs in the Umeå replication dataset. Only SNPs in the replication sample that reached a p-value of less than 0.05, and with the same direction of interaction, were considered to be truly replicated. According to these criteria, 51 SNPs from the previously identified pairs were replicated. There were no signs of statistically significant multiplicative interactions for any of these 102 smoking–SNPs pairs in either dataset, or for any of the remaining markers on the Immunochip.

Given extensive LD in the *HLA* region, we decided to adjust for HLA-SE alleles on the previously identified smoking–SNP pairs to determine possible interaction signals independent of HLA-SE. In EIRA, the numbers of interacting SNPs were strongly reduced after such adjustment with 12 being identified (after Bonferroni correction) from the co-dominant model and eight from the recessive model (five overlapping SNPs). Altogether eight non-duplicated SNPs (three from the co-dominant model and five from the recessive model: rs3104413, rs3129769, rs6931277, rs10484565, rs3129890, rs3129891, rs9268557 and rs9784858) were successfully replicated in the Umeå sample. All of these were from *HLA* class II region and were in high LD.

Furthermore, we analysed all the currently identified RA-risk SNPs, based on the published results, for interaction with smoking. We did not find any evidence for significant interactions

between these non-*HLA* SNPs and smoking on either the additive or the multiplicative scale in any of the three models, using a p-threshold based on adjustment for multiple comparisons.

Conclusions: Gene–smoking interactions were identified in ACPA-positive, but not ACPA-negative, RA. Notably, variants in *HLA-DRB1* and those in additional genes within the *MHC* class II region, but not in any other gene regions, showed interaction with smoking.

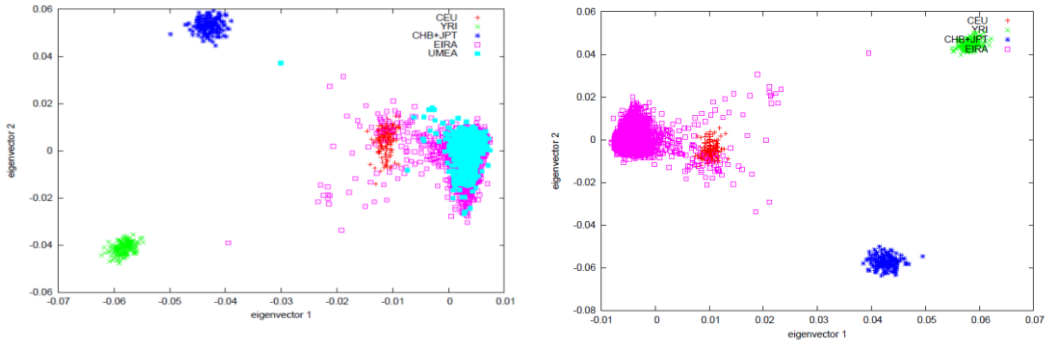


Figure5. Population stratification patterns of Umeå and EIRA samples.

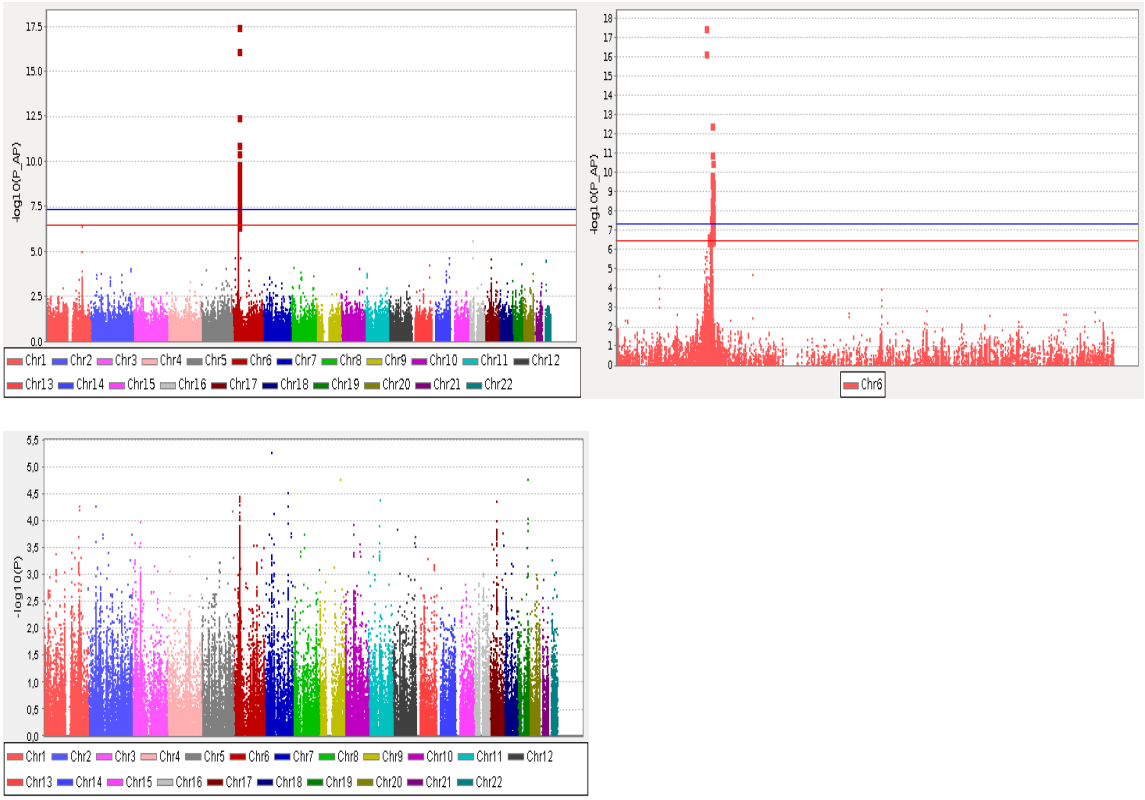


Figure6. Manhattan plot using p-values from the interaction analyses between smoking and Immunochip SNPs, for ACPA-positive RA (all chromosomes; upper left panel), ACPA-positive RA (chromosome 6; upper right panel) and ACPA-negative RA (lower panel).

5.2 STUDY II

Data were available from a total of 1998 cases and 2252 controls for the snuff–RA association analysis. In RA overall, 254 (12.9%) cases were ever moist snuff users and the number among controls was 290 (12.9%), resulting in an OR of 1.0 (95% CI: 0.8–1.2). We

did not observe any significant associations when the analysis was confined to ACPA-positive (OR: 1.0, 95% CI: 0.8–1.3) or ACPA-negative RA (OR: 0.9, 95% CI: 0.7–1.2). In addition, analyses based on current or past moist snuff use did not reveal any significant associations related to the risk of any RA subsets.

We then performed a stratified analysis on smoking status. Among never-smokers, no significant differences in RA risk were observed in the ACPA-positive subset (ever snuff users: OR: 1.0, 95% CI: 0.5–1.9; current snuff users: OR: 1.0, 95% CI: 0.5–2.2) as compared with the ACPA-negative subset (ever snuff users: OR: 1.0, 95% CI: 0.5–2.0; current snuff users: OR 1.5, 95% CI: 0.7–3.3). Of note, the numbers of observations were low, especially in the ACPA-negative group. Similarly, among ever-smokers, neither current nor former moist snuff use was associated with the risk of ACPA-positive or ACPA-negative RA.

Because the number of moist snuff users among never smokers was low (5.1%), we decided to increase the sample size by incorporating non-regular smokers and former smokers with long-term smoking cessation (women >12 years and men >32 years) into the never-smoker group. Again, we observed no significant association between current or former snuff use and the risk of any type of RA.

Finally, we performed the above-mentioned analyses among men and women separately. Essentially no differences were found, compared with results from both genders combined.

Conclusions: The use of moist snuff was not associated with the risk of either ACPA-positive or ACPA-negative RA, although a moderate protective or harmful effect could not be ruled out. The increased risk of RA associated with smoking is most probably not due to nicotine.

5.3 STUDY III

Smoking behaviour varied considerably by population, with the Korean cohort showing lower prevalence as compared with northern European populations. Heavy smoking (>10 pack-years), which is known to be a strong environmental risk factor for RA, was consistently associated with RA susceptibility in all three populations.

We imputed HLA classical alleles, amino acid residues and SNPs within the extended *MHC* region from the Immunochip data of 3588 EIRA, 598 NHS and 2125 Korean subjects. Concordance rates between the imputed and typed *HLA-DRB1* alleles were high (two-digit and four-digit resolution: EIRA, 97.3% and 95.0%; NHS, 93.8% and 91.8%; Korean cohort, 97.4% and 91.4%, respectively).

We assessed the associations between amino acid positions 11, 13, 71 and 74 in HLA-DRβ1, position 9 in HLA-B and position 9 in HLA-DPβ1 and the risk of seropositive RA. Consistent with previous reports, the most significant associations were found at amino acid positions 11 and 13, in all three populations. Relatively weak or no associations at the other positions in HLA-DRβ1, -B and -DPβ1 were identified.

The haplotype defined by amino acid positions 11, 13, 71 and 74 is the best model to explain the *HLA-DRB1* association with RA susceptibility. We subsequently found that RA was strongly associated with the dichotomised haplotype GRS in all three populations. When evaluating the interaction effect between the dichotomised haplotype GRS and heavy smoking, we found significant additive interactions in all three populations (AP (95% CI), p-value for AP: EIRA, 0.42 (0.31–0.53), 1.16×10^{-13}; NHS, 0.47 (0.19–0.75), 1.07×10^{-3}; Korean cohort, 0.80 (0.54–1.06), 9.73×10^{-14}).

We further evaluated the additive interaction between GRS at each amino acid positions and heavy smoking. Significant and consistent synergy was mapped to HLA-DRβ1 amino acid positions 11 and 13 in all three populations (AP 0.32–0.72, p-value 1.64×10^{-4} –1.45×10^{-5}). By contrast, there was no consistent interaction effect at the other positions. We also calculated multiplicative interaction effects and the results are summarised in Table6.

Conclusions: The significant gene–environment interaction effects indicate that there may be a physical interaction between citrullinated auto-antigens produced by smoking and HLA-DR molecules is characterised by the HLA-DRβ1 four-amino acid haplotype, primarily by the positions 11 and 13, in addition to the known SE positions 71 and 74.

Table6. Multiplicative interaction between amino acid GRS and heavy smoking in the risk of ACPA-positive RA.

Amino acid GRS	EIRA		NHS		Korea	
	OR* (95% CI)	p-value	OR* (95% CI)	p-value	OR* (95% CI)	p-value
GRS (binary)*pack-year (binary)						
Haplotype	1.09 (0.79–1.50)	0.62	1.77 (0.85–3.66)	0.12	1.77 (0.44–7.09)	0.42
P11	1.12 (0.79–1.58)	0.53	1.19 (0.89–1.58)	0.23	1.86 (0.68–5.14)	0.23
P13	1.15 (0.83–1.57)	0.40	1.19 (0.89–1.59)	0.25	2.14 (0.77–5.95)	0.14
P71	0.93 (0.69–1.24)	0.61	1.19 (0.77–1.82)	0.44	2.01 (0.63–6.41)	0.24
P74	1.19 (0.89–1.58)	0.24	0.79 (0.52–1.22)	0.29	1.43 (0.19–10.64)	0.73
P9-B	0.71 (0.50–1.00)	0.50	0.77 (0.35–1.69)	0.52	NA	0.27
P9-DPB1	1.07 (0.74–1.58)	0.73	0.34 (0.14–0.84)	0.02	NA	NA
GRS (continuous)*pack-year (continuous)						
Haplotype	1.01 (1.01–1.02)	<0.0001	1.00 (0.99–1.01)	0.53	1.03 (0.98–1.09)	0.30
P11	1.01 (1.00–1.01)	<0.0001	1.00 (0.99–1.01)	0.69	1.05 (1.00–1.11)	0.03
P13	1.01 (1.00–1.01)	<0.0001	1.00 (0.99–1.01)	0.80	1.04 (0.99–1.09)	0.05
P71	0.99 (0.99–1.00)	0.03	1.00 (0.99–1.01)	0.75	1.01 (0.97–1.06)	0.58
P74	1.01 (1.00–1.01)	<0.0001	0.99 (0.98–1.01)	0.27	1.08 (1.00–1.16)	0.05
P9-B	0.99 (0.98–1.01)	0.18	1.00 (0.98–1.02)	0.96	1.32 (0.70–2.49)	0.39
P9-DPB1	1.01 (1.00–1.02)	0.03	0.97 (0.95–1.00)	0.02	0.75 (0.56–1.01)	0.06

*: Odds ratios were adjusted for the top five PCs.

5.4 STUDY IV

A total of 6916 EIRA participants (1828 ACPA-positive cases, 1016 ACPA-negative cases and 4072 controls) were available for the study. Overall, 11.9% of ACPA-positive cases had at least one FDR with RA; the corresponding figures for ACPA-negative cases and controls

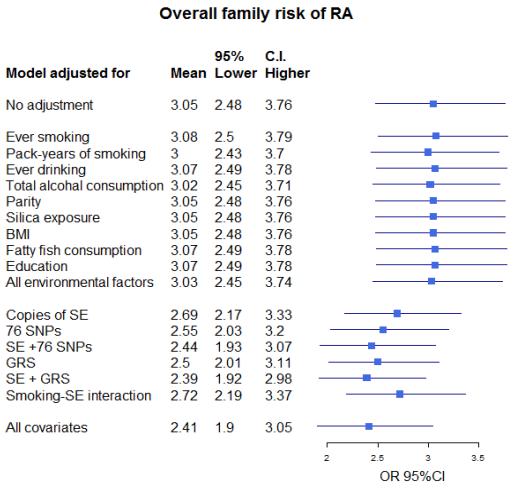
were 6.5% and 3.5%, respectively. We observed a higher proportion of FDRs with RA among control subjects who had donated blood samples (for genetic measurements) compared with those who had not; whereas the opposite pattern was observed among cases. To correct for this non-random distribution of missingness, MI was incorporated.

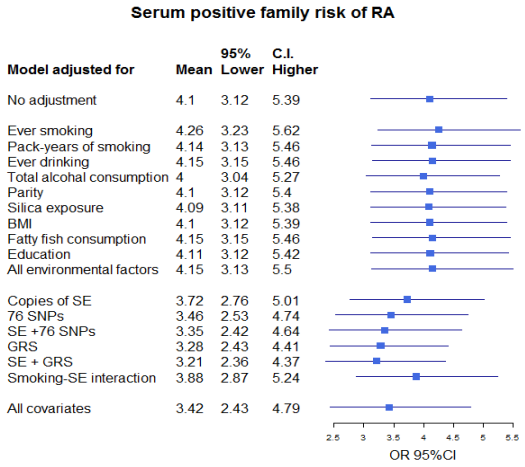
The crude familial OR for overall RA was 3.05 (95% CI: 2.48–3.76). When sequentially adjusting for each environmental risk factor, we observed no appreciable changes in the magnitude of familial risk, with ORs ranging from 3.00 to 3.08 (Figure7). However the estimates showed a moderate decrease after adjusting for genetic risk factors. Estimates from the full model, adjusting for all factors, were slightly lower than those adjusted for genetic factors alone; and adjusting for SE and smoking, and their interaction, did not further lower the familial risk as compared to adjusting for SE alone. The OR from the full model (2.41) was decreased by 21.1% based on using the beta coefficient (log(OR)) scale and by 31.2% using the OR scale, as compared with the OR from crude model (3.05).

When the analyses were separately performed according to ACPA status, we found a stronger familial aggregation in ACPA-positive RA (crude OR: 4.10) than in ACPA-negative RA (crude OR: 1.61). Although familial OR of ACPA-positive RA did not change considerably when environmental factors were included (adjusted OR range: 4.00–4.26), the value decreased to some extent when genetic risk factors were included (Figure7). In contrast to ACPA-positive RA, the familial OR of ACPA-negative RA was not influenced by either genetic or non-genetic factors. Adjusting for the literature-based GRS seems to explain a similar proportion of the familial aggregation as adjusting for the individual 76 SNPs for both ACPA-positive and ACPA-negative RA.

A complete-case analysis was also conducted. The estimates were in line with the imputed results in that environmental risk factors did not appear to explain the familial risk, while genetic factors only explained a small part for ACPA-positive RA.

Conclusions: Established risk factors only partly provide an explanation for the familial risk of RA, suggesting that many (familial) risk factors remain to be identified, in particular for seronegative RA. Family history therefore remains an important clinical risk factor for RA.





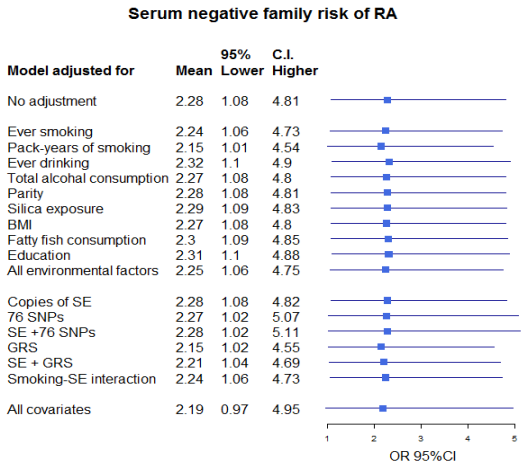


Figure7. OR and 95% CI of the familial risk in overall, ACPA-positive and ACPA-negative RA, by adjusting for each identified genetic and environmental risk factor.

6 DISCUSSION

Our results have confirmed previous established findings of the significant additive interaction effect between smoking and the *HLA* region, in the aetiology of RA. Furthermore, our results have revealed that the increased risk of RA associated with smoking is most probably not due to nicotine. With regard to the SE, the amino acid positions 11 and 13 appear to exert more effect than the traditional 71 and 74 SE positions. Finally, our results also indicate that the currently known genetic and environmental risk factors only account for a small proportion of RA familial risk, and additional factors need to be identified.

6.1 GENERAL METHODOLOGICAL CONCERNS

6.1.1 Power

Firstly, the "null findings" of the current results will be discussed from an epidemiological point of view. In Study I, we did not identify any other SNPs outside chromosome 6 that interacted with smoking in ACPA-positive RA. This might be due to a lack of power to determine small effect risk alleles. According to our detailed power calculation, using the current EIRA sample size and an alpha of 0.05, the power reaches 80% when the effect size of smoking is 1.5, the effect size of SNP is 1.1, and the synergistic effect size of smoking–SNP is 3.3; this might be true for *HLA* alleles with a strong impact in RA, but may not always be the case for other alleles outside *HLA* with relatively weak magnitudes. Furthermore, in a genome-wide setting in which multiple comparisons have been made, it is challenging to retain this high level of power when the true alpha needs to be corrected based on the numbers of tests performed. Similarly, this may explain the negative findings in ACPA-negative RA as this subgroup only constitutes 30% of the total patient population, and is believed to be heterogeneous rather than homogenous, making it even more likely to be underpowered. In Study II, we have estimated the effect of snuff use, with an exposure rate of 5%, to be around 1.0; this effect size is so small that to rule out a moderate protective (or harmful) influence of snuff on RA risk in the order of 1.17 (with an alpha level of 0.05), 10000 cases and 10000 controls would be needed.

There are several ways to boost power in epidemiological studies in genome-wide settings.[80,148-150] The most straightforward strategy is to increase sample size, which can be implemented by either the involvement of more controls or the combination of datasets from several different populations to perform a meta-analysis. The marker imputation technique and population stratification analysis have made this strategy even more practical nowadays. Where different cohorts have used different genotyping chips, imputation can be applied to equate the set of SNPs in each study resulting in a "full" chip.[80] Subsequently the studies can be combined for each SNP. An alternative way to increase power is to incorporate a multi-stage design. Because the significance criteria for GWAS are usually quite strict, with corrections for multiple comparisons adopted to avoid false-positive findings, power is therefore much less than might be imagined. A multi-stage design employs a first stage with a small sample size and a low significance level to detect all possible loci of realistic effect

size. These small fractions of markers that pass the first stage will be subsequently evaluated in an independent sample at the second stage, which is similar in size to or larger than the first population. A large efficiency will thus be gained due to the considerably reduced number of markers in the second stage.

Despite our attempts to increase power for Studies I, III and IV, through a two-stage design, a combined meta-analysis, allele imputation and missing data imputation, the issue still remains a major concern in studies of this type, where hundreds of thousands of markers with weak effects are investigated. A better way to improve the insufficiency of power is to collaborate universally with the inclusion of all available datasets. The utilisation of consortium data would be a reasonable next step.

6.1.2 Bias

Bias has been classified traditionally into three broad categories, selection bias, information bias and confounding. Selection bias occurs where the exposure frequency does not reflect that of the study base. For example, in a screening test, the study subjects usually volunteer to be tested, i.e. they select themselves to be screened, whereas the non-participants choose not to be screened, thus a selection bias could occur. Recall bias indicates a different pattern with regard to the accuracy of information collected, between cases and controls. For example, the patients are more likely (or they try) to remember exposures more often (or correctly) (e.g. smoking, which might be considered by them as an important RA risk factor) concerning their disease compared to the control subjects. Because this over-recall or over-report is related to the disease, it tends to result in differential misclassification.[151] The EIRA study has several strengths that reduce bias to a large extent. Firstly, the universal free access to the medical care system provided in Sweden makes it less likely that people would avoid seeking medical help due to financial concerns. Therefore, a relatively complete set of representative patient data could be captured. Within the study area, all public as well as most privately run rheumatology units were linked to the general welfare system, reporting cases to the EIRA database. Its population-based design recruiting incident RA cases, where the estimated median duration from first symptom onset to disease diagnosis was 195 days and the estimated time between diagnosis and completing the questionnaire was within 12 months, makes recall bias less likely than for other study designs (i.e. study using prevalent cases). Moreover, the newly diagnosed cases derived from the population share very different traits as compared with cases recruited from hospitals (hospital-treated patients), where the latter subset might be older, have a more advanced disease course and worse in prognosis, thus the likelihood of selection bias is decreased. Similarly, in EIRA, controls were randomly selected from a nationwide register, reflecting the characteristics of the study population. Additionally, the high participation rates in EIRA (more than 90% of the invited cases and more than 75% of the invited controls) further minimised bias.

However, despite the above-mentioned strengths, we have indeed observed some biased behaviour among cases and controls in terms of donating blood. In Study IV, control subjects who provided blood samples were more likely to have a family history of RA, whereas the

opposite was true for cases. Because analysis of genetic data could only be performed among participants who had donated blood samples, such non-random missingness should always be taken into account. In Studies I, II and III, we did not observe any differences in the distributions of smoking or snuff use among those subjects who did and did not donate blood samples.

Confounding refers to factors that are associated with both the disease and the exposure, yet are not an effect of the exposure. In Studies I and III, in which genetic factors have been examined, fewer confounding factors need to be considered, as the exposure itself is very unlikely to be influenced by other factors. In Study II, in which we analysed the association between snuff use and RA, smoking becomes the most important confounder with the largest effect. Two methods have been implemented to control for confounding. Firstly, EIRA cases and controls were matched by age, gender and residential area, the most common potential confounders. Secondly, the sample size of EIRA allowed us to perform a stratified analysis, restricted to non-smokers. We have tried to adjust for various other environmental risk factors and the results were not affected considerably as compared with adjusting only for matching factors and smoking. Of note, although the aim of Study IV is to determine to what extent RA familial risk could be explained by current risk factors, the methods used to reach this aim are the same as those used to control for confounders. We hypothesised that the familial risk has no direct influence on the patients' outcome, but rather exerts an effect through a number of shared genetic and environmental factors. The well-established EIRA questionnaire, covering a wide range of lifestyle-related questions, as well as the genotyping data from blood samples, makes it possible to collect data on all the currently known RA risk factors and adjust for their effects. After excluding the effects of these "confounders", we expected to see the true remaining familial risk of RA. This was in fact substantial, indicating the major role of familial history in RA risk, and suggesting that there are more factors to be identified (provided the results were not due to residual confounding from the factors for which we adjusted).

6.1.3 Treatment of Missing Data

Missing data could be commonly classified as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).[152] MCAR is the easiest to understand: in a dataset, any part of the data is equally likely to be missing as any other part, and no relationship is likely between the missing and observed values. The difference between MAR and MNAR lies largely in whether the systematic difference between the missing values and the observed values can be explained by differences in the observed data. For example, depressed people might be less willing to report their incomes, and also have a lower income in general. Thus although a high proportion of missing data among depressed individuals could be observed, the missingness would be unrelated to income level but rather related to their depression; this is an example of MAR. On the other hand, people with a low income are less likely to reveal their income; this is an example of MNAR. Unfortunately, in reality, it is not possible to distinguish between MAR and MNAR, and sometimes it can also

be challenging to recognise MCAR. Therefore, previously a variety of approaches have been frequently used to deal with missing data, including replacing missing values with values imputed from the observed data (for example, use of the mean of the observed values to replace missing data), adopting a missing category indicator and replacing missing values with the last measured value (last value carried forward).[153] None of these approaches is statistically valid or sound in general, as they are based on models with implausible assumptions. Moreover, these methods impute the missing data only once and then proceed to the completed data analysis, which can lead to serious bias, and are thus rarely recommended.

An arbitrary way of dealing with missing values is to simply exclude subjects without values. This method was used in Studies I, II and III in which only limited environmental variables (smoking or snuff) were used as the main covariates, each with virtually complete information. We included subjects with full information on both genotyping and environmental data in our analyses. The disadvantage of this method is that we may lose power, but the estimates remain most probably unbiased by the absence of data. However, in Study IV, in which a wide variety of environmental as well as genetic factors have been examined, implementing complete-case analysis would reduce the sample size to half of the original size, and thus result in a great loss of precision and power. We therefore employed the MI method, which allows individuals with previous incomplete data to be included by assigning an imputed value. The imputed values are generated on the basis of existing data, irrespective of genetic or environmental factors, based on a Bayesian approach.[154] MI accommodates the uncertainty about the missing data by generating several different plausible imputed datasets and appropriately combining results obtained from each of them. Notably, MI is usually based on the assumptions of MAR and normally distributed data. MI might be a superior method of dealing with missing values compared with previous approaches that lack plausible assumptions. However, taking into account the MI modelling, it might not be surprising that similar patterns of covariate distribution between observed and imputed data could be identified after MI, and that given this similar distribution pattern, the estimates based on imputed data would be in line with estimates based on complete-case analysis. Other practical ways of dealing with missing data include non-response weighting and likelihood-based methods, which could provide a good solution to this problem.[155]

6.2 FINDINGS AND IMPLICATIONS

6.2.1 *HLA* Remains an Important Genetic Region in RA Aetiology

The heritability of RA was firstly (and classically) evaluated through twin studies; the increased prevalence of a trait in monozygotic (MZ) twins (who share 100% of their genome sequence) compared to dizygotic (DZ) twins (who share 50% of their genome sequence) is a powerful way to measure the importance of genetics in the variation of that trait. The concordance rate for RA in MZ twins is around 12–15%, approximately four times greater than in DZ twins, which indicate that a considerable genetic contribution in RA.[41] These concordance rate data, after being transferred by using quantitative genetic methods, suggest

that RA heritability is around 40–60%.[41] The estimates have been validated in a recent large twin study, showing that ACPA-positive RA heritability is around 39–44%.[85] By determining MZ twin concordance rates, together with the empirical sibling recurrence risk and the observed *HLA* haplotype sharing by pairs of affected siblings, the contribution of the *HLA* region to RA heritability has been estimated to be 37%.[58] Consistently, Recent studies using genome-wide markers have estimated the identified *HLA* region variants would explain 25.4% of ACPA-positive RA heritability.[86,87] These data together with the results from Study I provide clear evidence that the *HLA* region continues to be an area of major interest in RA aetiological studies.

Extensive investigations of the association between the *HLA* region and RA have been performed, in an effort to clarify the complex hierarchy of risk factors conferred by different *HLA* genotypes. Association studies have demonstrated that, in European populations, all *HLA-DRB* alleles with the SE (01, 04 and 10) provide RA-prone antigen recognition, increase the risk of developing ACPA-positive RA and extra-articular manifestations, increase the likelihood of progressing into a more severe, erosive, deforming disease and are responsible for poor prognosis.[156] A gene–dose effect could be observed, which is compatible with the role of *HLA* polymorphisms in T cell repertoire shaping.[157] There has been some debate regarding the role of *DRB1*15*, with a few studies finding a linkage between *15 and enhanced ACPA production or circulation.[158,159] By contrast, *DRB1*13* has been shown to exert a protective effect.[160] Some other alleles that are negatively associated with RA include *0103, *0402 and *0802. Haplotype analysis revealed that *DQ* had an important modifying influence on the risk of individual SE alleles, resulting in greater disease severity, RF positivity and greater degrees of joint deformity. The results of some association studies have suggested a direct role for *DQ* alleles in RA whereas further larger studies have not supported this hypothesis.[99] Conflicting findings have been reported with regard to other susceptible loci within the same region independent of SE, including *A1-B8-DR3, ZNF311, TNF, DP, DO TAP, MICA, VARS2L* and others.[156]

It is clear why some *HLA* associations, despite having been extensively studied, remain controversial. One very important consideration is ethnic or racial differences; another major influence comes from LD, a problem that is not unique to RA but affects studies of all diseases with strong *HLA* associations. Several solutions to overcome the strong LD in this region have been suggested. A common approach has been to match cases and controls for the haplotypes at *HLA-DRB1*, and to use large datasets to obtain sufficient power. Other solutions include using the within-family association, or pooling on the basis of carriage of a specific *DRB1* allele.

Before 2012, *HLA* alleles were believed to be exclusively associated with ACPA-positive RA. Then, a well-powered study combining data from several Caucasian populations identified and confirmed the association between SE and ACPA-negative RA.[161] Despite adjustment for the heterogeneity of ACPA-negative RA in the study, as well as validation using clinically homogeneous ACPA-negative cases, i.e. CCP-negative RA cases which were

also negative for four different ACPAs (α-enolase, vimentin, fibrinogen and collagen type II), the role of this identified association and its possible restriction to specific subtypes of ACPA-negative RA remain to be determined. It also remains a challenge to identify interactions or epistasis among ACPA-negative disease subsets, firstly because ACPA-negative RA is likely to be a mixture of arthritis-related symptoms rather than a homogenous disease group; and secondly because the sample size is far smaller for ACPA-negative compared to ACPA-positive RA cases; and finally, so far, no strong or consistent environmental or genetic risk factors have been found in ACPA-negative RA, making it even more difficult to identify interaction effects without any main association effects. Therefore, further effort is needed to collect large numbers of "pure" ACPA-negative patients, and to identify pathogenically relevant subsets within this population of RA patients.

6.2.2 Interactions outside the HLA region Remain to be Identified

Despite the strong linkage between *HL*A and RA, the presence of *HLA* alleles is neither necessary nor sufficient for occurrence of the disease. The remaining risk could be ascribed to other regions. Moreover, previously published data using the candidate gene approach have shown that smoking interacts with genes from other chromosomes, with one well-established example being *PTPN22*.[142] A possible way to identify the potential signals could be through incorporating biological mechanisms to connect specific genes or gene pathways to specific environmental factors. The interaction between smoking and the *HLA* region is compatible with the arthritogenic antigen-presentation theory. Alcohol consumption might exert its effects through alternative gene pathways.[162] For example, results from studies in mice suggest that: 1) levels of tumour necrosis factor and interleukin-6 (two pro-inflammatory molecules implicated in RA) can be reduced by adding a low dose of ethanol to the drinking water; 2) NSAIDs cause gastric/gastrointestinal pain and bleeding while DMARDs worsen liver problems, and both effects can be exacerbated by alcohol; and 3) the cytokine–hormone axis might be another source of genetic pathways when considering the alcohol–gene interaction. More gene–gene and gene–environment interactions remain to be revealed, although they might be weaker in magnitude than the smoking–SE interaction. One very interesting possibility could be to explore the roles of some well-recognised inhalation factors, and their synergistic effect with genes, as these factors share the same exposure pathways with smoking: the airways and the lung. Examples of such factors could include textile dust, silica dust, air pollutants and solvents.

6.2.3 Smoking Is a Major Preventable Factor for RA

Smoking is a well-characterised inhalation exposure, and the smoking–RA association is the most recognised link between the environment and the aetiology of the disease. A large number of studies have demonstrated adverse effects of smoking in either RA incidence or prognosis.[105 106] In line with previous findings, we confirmed a comparable risk effect of smoking in ACPA-positive RA among one Asian and three European populations from Studies I and III. In Study II we also assessed the influence of smoking cessation on RA risk; consistent with previous findings, the effect of smoking started to return to baseline after 12

years of cessation among women and after 32 years among men, indicating a reduction but not elimination of the risk. The excess fraction (EF) of cases attributable to smoking has been calculated as an indicator of the relevance of smoking as a risk factor for RA in the population of Sweden. It was concluded that EF attributable to smoking was 35% for ACPA-positive RA and 20% for RA overall; in addition, among ACPA-positive RA cases with double SE alleles, 55% could be attributable to smoking.[10] Given that EF is highly dependent on the prevalence of exposure, it may be higher in other populations with higher smoking rates than in Sweden, a country with a low prevalence of smoking.

From a public health perspective, the additive interaction between smoking and genes provides optimal information in terms of disease prevention and intervention: if the joint effect of two factors is higher than the sum of their single effects, then reduction of either factor would also reduce the risk of the other in producing disease.[163] Taking into account the profound synergistic effect, as well as that it takes more than 10 years for the main effect of smoking to return to the baseline level, it is important to advise RA patient not to start smoking, to smoke less or to quit as soon as possible. A more sensible practical strategy could be to educate the families, and in particular the children, of RA patients about the importance of not smoking.

In Study II, we found no association between snuff use and the risk of RA. We conclude that constituents in the cigarette smoke other than nicotine, most probably many noxious substances that may cause irritation of the airways and activation of innate as well as adaptive immunity, are likely to be involved in the pathogenesis of RA. However, we could not rule out a minor effect (harmful or protective) of nicotine in RA, and large studies are warranted to elucidate its association with RA. However, we do not recommend that RA patients use snuff as a substitute for cigarette smoking, because snuff is a demonstrated risk factor for oral cancer.

6.2.4 Reconsideration of the Definition of SE

The SE hypothesis was first proposed by Gregersen *et al.* almost 30 years ago,[43] which may no longer be the best or the most complete model to describe RA risk in light of the dramatic changes in technology and biology that have since occurred. Attempts have been made to redefine SE. In 2005, Sophie Du Montcel *et al.* proposed a new classification of *HLA-DRB1* alleles,[164] according to which, the risk of developing RA depends on whether the RAA epitope is present at positions 72–74 but is also modulated by the amino acids in positions 70 and 71. The KRAA motif at positions 71–74 confers the highest RA susceptibility, and the RRRAA or QRRAA motifs confer an intermediate risk.[164] This new classification was subsequently tested and validated by Laetitia Michou *et al.*, in an independent sample of 100 Caucasian RA trio families,[165] as well as by Thomas Barnetche *et al.*, in 759 cases and 789 controls with different ethnic backgrounds.[166] However, this new classification was restricted to traditional SE with no novel positions involved. In 2012, Raychaudhuri *et al.* applied an imputation approach to SNP data from a GWAS meta-analysis in 5018 seropositive RA cases and 14974 controls, and demonstrated that the risk of RA associated with *HLA-DRB1* gene

correlates most strongly with the amino acid residue in position 11 (or 13) located at the bottom of the surface of the DRβ1 antigen-binding groove.[56] The traditional SE positions 71 and 74, were also but not as strongly associated independently with RA susceptibility. In addition, independent RA risk alleles in *HLA-B* and *HLA-DPB1* were found. No further association signals were identified within the *HLA* when controlled for all these independent effects from the five amino acid positions. Results from our Study III support and strengthen the finding of Raychaudhuri *et al.* that the most profound interaction effect with smoking was on the amino acid position 11 or 13, in addition to the amino acid positions 70–74, and that the haplotype based on amino acids 11, 13, 71 and 74 had more pathogenic effect in the binding and presentation of smoking-induced auto-antigen like ACPAs. Based on these results, the *HLA-DRB1* association and the SE-smoking interaction appear to be best explained by the new haplotype, but biological explanations are still needed.

6.2.5 Uncharacterised Genetic Variance Remains to be Discovered

The results from Study IV suggest that all the currently identified RA risk factors together only explain a small proportion of the total susceptibility. It has been estimated that hundreds of common risk alleles are likely to exist but remain undiscovered to date, and that those uncharacterised SNP associations throughout the genome, together with known risk alleles, would explain in total 36% of RA disease risk.[87] Therefore, current SNP associations only account for half of the estimated RA heritability. Sequencing experiments might have the potential to identify causal variants across the entire allele frequency range, in particular for low frequency alleles. A recent epigenome-wide association study in 354 ACPA-positive RA patients and 337 control subjects identified 10 differentially methylated positions potentially mediating genetic risk in RA, nine of which were located within MHC over four gene regions and the only one outside of MHC was located at chromosome 6.[167] These findings, together with our results, further indicate a great potential for the identification of genetic or epigenetic variations outside of the MHC; within the MHC, a challenging task for future investigations will be to determine which specific immune reactions are related to smoking and specific MHC molecular structures.

6.3 FUTURE DIRECTIONS

- Further identification of causal genetic variants and their characterisation, and incoporation of epigenetic results. Development of computational tools to calculate genome-wide interactions taking into account all the possible combinations of current markers available from the CHIP (not only one-to-one interaction, but also N-to-N interaction).
- Investigation of gene–gene and gene–environment interactions involving environmental risk factors other than smoking, such as lifestyle-related factors (e.g. alchol consumption, body mass index and others). In particular, inhalable factors (e.g. silica dust, solvent exposure and textile dust) and airway infections (e.g. influzena) should be considered.
- Accurate risk prediction in susceptible individuals through the combination of currently identified genetic markers, lifestyle-related factors, occupational hazards and serological measurements. Accurate prediction of disease heritability in susceptible individuals using their familial risks and genotypes, and information from relatives.
- Utilisation of consortium data and collaborative efforts involving multiple populations to increase the power of current research.

7 ACKNOWLEDGEMENTS

First of all, I wish to thank all the **EIRA participants**, both **cases** and **controls**, for their contributions to our study and for making my doctoral thesis possible. I would like to thank all EIRA administrative personnel: our research assistants **Lena Nise**, **Marie-Louise Serra**, **Caroline Öfverberg**, **Ida Gärskog**, **Edit Ekström** and **Anna Peterson**, project coordinator **Charlotte Anderberg** and database managers **Ville Söderberg** and **Diogo del Gaudlo**, for your excellent assistance with the collection and management of data, which has made the projects run smoothly over the years.

My PhD is funded partly by the Chinese Scholarship Council (CSC) and partly by the Institute of Environmental Medicine (IMM). I wish to express my sincere gratitude to the **CSC** and **IMM** for covering the salary, especially **HR personnel and accountants at IMM** for calculating the salary for me.

During my 4-year PhD period, I have had the opportunity and pleasure of working with brilliant and dedicated colleagues, many of whom I consider best friends. In particular, I would like to acknowledge the following people.

Lars Alfredsson, my main supervisor, as well as the head of the Cardiovascular Epidemiology Unit. Thank you for your trust and support, for offering me the opportunity of working with excellent collaborators worldwide, for sharing epidemiological knowledge with me, for many uplifting talks, for all the nice scientific discussions we have had. Thank you for your positive and realistic attitude towards everything which has inspired me to be more of a "doer".

Lars Klareskog, my co-supervisor, as well as the head of the rheumatology unit. Thank you for sharing your broad knowledge of immunology with me, for depicting the "blueprint" for the whole field of rheumatology for us, for guiding me through tough projects through frequent meetings and interesting discussions, for your constructive and sincere advice on my career plan.

Henrik Källberg, my co-supervisor. Thank you for always being supportive and working with me on programming, data analysis and literature reading. Thank you for many uplifting talks, for always being considerate to me, for exchanging ideas and thoughts.

Leonid Padyukov, my co-supervisor. Thank you for sharing your broad knowledge of biology: my first discussion was with you sitting in the corridor of the rheumatology unit, recalling all the concepts regarding genes, SNPs, base-pairs, allele frequencies etc. Thank you for many useful suggestions.

Camilla Bengtsson, my first project was a collaboration with you. We then collaborated on another three projects although you were never my official supervisor. Thank you for the nice time working together with great efficiency, for getting ideas and inspiration from each other,

for your excellent supervision, for your remarkable writing skills, for your broad knowledge of the EIRA data.

My collaborators from the Karolinska Institutet: **Thomas Frisell**, **Johan Askling**, **Karin Lundberg**, **Nastya Kharlamova**, **Evan Reed**, **Maria Sandberg** and **Saedis Saevarsdottir** we have worked together to certain extents and I have learned greatly from you. Thank you for trusting me in handling the data, and for your professional contributions to our papers.

My collaborators abroad: **Elizabeth W. Karlson**, **Jing Cui**, **Chia-Yen Chen**, **Kwangwoo Kim** and **Jeffrey Sparks** at the Harvard School of Public Health and Harvard Brigham and Women's Hospital, thank you for those extremely productive and efficient days under your "training". **Diane van de Woude** and **Leendert Trouw** at Leiden University, thank you for the great discussions over the telephone and by email. **Mingfen Ho** and **Tim Bongartz** at the Mayo Clinic, thank you for your efforts to make our collaboration finally possible.

All the office roommates I have had over the years: **Anna Ilar**, **Zuomei Chen**, **Dashti Dzayee**, **Cecilia Orellana Pozo**, **Ilais Moreno**, **Gholamreza Abdoli** and **Zahra Golabkesh**, you are friends and colleagues of great importance to me. Thank you for all the encouragement and support to lift me up, for reaching out without any hesitation when I encountered difficulties, for the nice learning and hanging out time together, for the birthday and Christmas cards and gifts, for inviting me home, for being so kind, understanding and helpful to me.

All the colleagues in our corridor: **Hedley Quintana**, **Hozan Hussen**, **Anette Linnersjö**, **Maria Bäärnhielm**, **Mohammad Mohammadi**, **Mohsen Besharat Pour**, **Bahareh Rasouli**, **German Carrasquilla**, **Federica Laguzzi** and **Paolo Frumento**, we started working at IMM at approximately the same time, and have shared a lot of feelings and experiences of research life, thank you for being there for me. **Max Vikström**, thank you for your consistent efforts in saving and managing all the printing papers, and for your great sense of humour. **Boel Brynedal** and **Helga Westerlind**, for organising the genetics-genomics journal club. **Karin Leander** and **Anita Berglund**, for guiding us through the epi-courses, and introducing me to those interesting seminars.

All my professors and classmates at the **National Clinical Research School in Chronic Inflammatory Disease**, most of whom work at CMM. Two years of attempts, extra courses, workshops and journal clubs, Class 2013, we made it!! Thank you especially **Radha Thyagarajan** and **Johanna Estelius**, for being there for me unconditionally, and for supporting me. We now have time to travel together more often; thank you for still inviting me, after several refusals because I was too occupied by work and study! We definitely should hang out more often.

All my Chinese friends: **Lidi Xu** and **Peng Zhang**, for accompanying me. **Chao Sun**, for your help. **Linjing Zhu**, **Zhangsen Huang**, **Chuxiao Huang**, for the memorable time with your family. **Jingli Yang** and **Jiayao Lei**, for exchanging thoughts and feelings with me, for

composing a “core friends” group for me and protecting me. **Jiangnan Luo** and **Chengjun Sun**, for organising the party and cooking for all of us. **Xiaolu Zhang** and **Bingnan Li**, for your company and kindness.

All the lovely kids who showed up from time to time in our office: **Sebastián**, **Sebastian**, **Ellen**, **Tore**, **Aston**, **Alicia** and **Daniel**, **Vilmer** and **Valter**, **Erasmus**, **Filippa**, **Vigo** and **Bruno**, it was great fun talking and playing with you, even for a short time.

Finally, I would like to express my appreciation to my parents, **Jianshu Dai** and **Zhonghong Jiang**, for raising me in a family full of warmth, harmony and fun, for loving each other over 30 years, for your positive and optimistic attitude towards life, for loving me unconditionally. Thank you to my cousins, **Xinli Dai** and **Kanyi Zhu**, you turned out to be very supportive during my toughest time in Sweden, kinship really means a lot. My grandfather, **Zidong Jiang**, hope you are happy and proud of me in heaven. My grandmother, **Huiming Sun**, thank you for being so healthy and for your longevity: a precious gift to our big family of four generations. My maternal grandparents, **Atu Dai** and **Meizhen Xu**, thank you for offering me all the good things during each of my visits, for making me a lot of new clothes through your work as professional tailors. Love you all!

8 REFERENCES

1. Klareskog, L., Catrina, A.I. & Paget, S. Rheumatoid arthritis. *Lancet* **373**, 659-672 (2009).

2. Lee, D.M. & Weinblatt, M.E. Rheumatoid arthritis. *Lancet* **358**, 903-911 (2001).

3. Scott, D.L., Wolfe, F. & Huizinga, T.W. Rheumatoid arthritis. *Lancet* **376**, 1094-1108 (2010).

4. Holoshitz, J. The rheumatoid arthritis HLA-DRB1 shared epitope. *Curr Opin Rheumatol* **22**, 293-298 (2010).

5. MacGregor, A.J., *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum* **43**, 30-37 (2000).

6. Frisell, T., *et al.* Familial risks and heritability of rheumatoid arthritis: role of rheumatoid factor/anti-citrullinated protein antibody status, number and type of affected relatives, sex, and age. *Arthritis Rheum* **65**, 2773-2782 (2013).

7. Costenbader, K.H., Feskanich, D., Mandl, L.A. & Karlson, E.W. Smoking intensity, duration, and cessation, and the risk of rheumatoid arthritis in women. *Am J Med* **119**, 503 e501-509 (2006).

8. Criswell, L.A., *et al.* Cigarette smoking and the risk of rheumatoid arthritis among postmenopausal women: results from the Iowa Women's Health Study. *Am J Med* **112**, 465-471 (2002).

9. Di Giuseppe, D., Orsini, N., Alfredsson, L., Askling, J. & Wolk, A. Cigarette smoking and smoking cessation in relation to risk of rheumatoid arthritis in women. *Arthritis Res Ther* **15**, R56 (2013).

10. Kallberg, H., *et al.* Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann Rheum Dis* **70**, 508-511 (2011).

11. Stolt, P., *et al.* Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case-control study, using incident cases. *Ann Rheum Dis* **62**, 835-841 (2003).

12. Klareskog, L., Padyukov, L., Ronnelid, J. & Alfredsson, L. Genes, environment and immunity in the development of rheumatoid arthritis. *Curr Opin Immunol* **18**, 650-655 (2006).

13. Klareskog, L., *et al.* A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis Rheum* **54**, 38-46 (2006).

14. Padyukov, L., Silva, C., Stolt, P., Alfredsson, L. & Klareskog, L. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis Rheum* **50**, 3085-3092 (2004).

15. Sopori, M.L. & Kozak, W. Immunomodulatory effects of cigarette smoke. *J Neuroimmunol* **83**, 148-156 (1998).

16. Abbas, A.K. & Lichtman, A.H. *Basic immunology : functions and disorders of the immune system*, (Saunders/Elsevier, Philadelphia, PA, 2009).

17. Neovius, M., Simard, J.F. & Askling, J. Nationwide prevalence of rheumatoid arthritis and penetration of disease-modifying drugs in Sweden. *Ann Rheum Dis* **70**, 624-629 (2011).

18. Majithia, V. & Geraci, S.A. Rheumatoid arthritis: diagnosis and management. *Am J Med* **120**, 936-939 (2007).

19. Eriksson, J.K., *et al.* Incidence of rheumatoid arthritis in Sweden: a nationwide population-based assessment of incidence, its determinants, and treatment penetration. *Arthritis Care Res (Hoboken)* **65**, 870-878 (2013).

20. Kelly, C. & Hamilton, J. What kills patients with rheumatoid arthritis? *Rheumatology (Oxford)* **46**, 183-184 (2007).

21. Listing, J., *et al.* Mortality in rheumatoid arthritis: the impact of disease activity, treatment with glucocorticoids, TNFalpha inhibitors and rituximab. *Ann Rheum Dis* (2013).

22. van der Helm-van Mil, A.H. & Huizinga, T.W. Advances in the genetics of rheumatoid arthritis point to subclassification into distinct disease subsets. *Arthritis Res Ther* **10**, 205 (2008).

23. Arnett, F.C., *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* **31**, 315-324 (1988).

24. Aletaha, D., *et al.* 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* **69**, 1580-1588 (2010).

25. Vonkeman, H.E. & van de Laar, M.A. The new European League Against Rheumatism/American College of Rheumatology diagnostic criteria for rheumatoid arthritis: how are they performing? *Curr Opin Rheumatol* **25**, 354-359 (2013).

26. Rhodes, B. & Vyse, T.J. General aspects of the genetics of SLE. *Autoimmunity* **40**, 550-559 (2007).

27. Panayi, G.S. Developments in the immunology of rheumatoid arthritis, a personal perspective. *Rheumatology (Oxford)* **50**, 815-817 (2011).

28. Scrivo, R., Di Franco, M., Spadaro, A. & Valesini, G. The immunology of rheumatoid arthritis. *Ann N Y Acad Sci* **1108**, 312-322 (2007).

29. Firestein, G.S. Evolving concepts of rheumatoid arthritis. *Nature* **423**, 356-361 (2003).

30. Raza, K. & Gerlag, D.M. Preclinical inflammatory rheumatic diseases: an overview and relevant nomenclature. *Rheum Dis Clin North Am* **40**, 569-580 (2014).

31. Jutley, G., Raza, K. & Buckley, C.D. New pathogenic insights into rheumatoid arthritis. *Curr Opin Rheumatol* **27**, 249-255 (2015).

32. Lewin, B., Krebs, J.E., Kilpatrick, S.T. & Goldstein, E.S. *Lewin's genes X*, (Jones and Bartlett, Sudbury, Mass., 2011).

33. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).

34. Glazier, A.M., Nadeau, J.H. & Aitman, T.J. Finding genes that underlie complex traits. *Science* **298**, 2345-2349 (2002).

35. Peltonen, L. & McKusick, V.A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291**, 1224-1229 (2001).

36. Klein, J. George Snell's first foray into the unexplored territory of the major histocompatibility complex. *Genetics* **159**, 435-439 (2001).

37. Carosella, E.D. From MAC to HLA: Professor Jean Dausset, the pioneer. *Hum Immunol* **70**, 661-662 (2009).

38. Klein, J. & Sato, A. The HLA system. First of two parts. *N Engl J Med* **343**, 702-709 (2000).

39. Muers, M. Complex disease: ups and downs at the MHC. *Nat Rev Genet* **12**, 456-457 (2011).

40. Neefjes, J., Jongsma, M.L., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* **11**, 823-836 (2011).

41. Newton, J.L., Harney, S.M., Wordsworth, B.P. & Brown, M.A. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun* **5**, 151-157 (2004).

42. Stastny, P. Mixed lymphocyte cultures in rheumatoid arthritis. *J Clin Invest* **57**, 1148-1157 (1976).

43. Gregersen, P.K., Silver, J. & Winchester, R.J. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum* **30**, 1205-1213 (1987).

44. Auger, I., *et al.* Influence of HLA-DR genes on the production of rheumatoid arthritis-specific autoantibodies to citrullinated fibrinogen. *Arthritis Rheum* **52**, 3424-3432 (2005).

45. Chun-Lai, T., *et al.* Shared epitope alleles remain a risk factor for anti-citrullinated proteins antibody (ACPA)--positive rheumatoid arthritis in three Asian ethnic groups. *PLoS One* **6**, e21069 (2011).

46. Hughes, L.B., *et al.* The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture. *Arthritis Rheum* **58**, 349-358 (2008).

47. Huizinga, T.W., *et al.* Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis Rheum* **52**, 3433-3438 (2005).

48. Kazkaz, L., *et al.* Rheumatoid arthritis and genetic markers in Syrian and French populations: different effect of the shared epitope. *Ann Rheum Dis* **66**, 195-201 (2007).

49. Marotte, H., *et al.* The shared epitope is a marker of severity associated with selection for, but not with response to, infliximab in a large rheumatoid arthritis population. *Ann Rheum Dis* **65**, 342-347 (2006).

50. Ding, B., *et al.* Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis

in the extended major histocompatibility complex region. *Arthritis Rheum* **60**, 30-38 (2009).

51. Lee, H.S., *et al.* Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the DRB1 locus. *Mol Med* **14**, 293-300 (2008).

52. Nordang, G.B., *et al.* HLA-C alleles confer risk for anti-citrullinated peptide antibody-positive rheumatoid arthritis independent of HLA-DRB1 alleles. *Rheumatology (Oxford)* **52**, 1973-1982 (2013).

53. Okada, Y., *et al.* Contribution of a haplotype in the HLA region to anti-cyclic citrullinated peptide antibody positivity in rheumatoid arthritis, independently of HLA-DRB1. *Arthritis Rheum* **60**, 3582-3590 (2009).

54. Orozco, G., *et al.* HLA-DPB1-COL11A2 and three additional xMHC loci are independently associated with RA in a UK cohort. *Genes Immun* **12**, 169-175 (2011).

55. Vignal, C., *et al.* Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. *Arthritis Rheum* **60**, 53-62 (2009).

56. Raychaudhuri, S., *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-296 (2012).

57. Okada, Y., *et al.* Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *Hum Mol Genet* (2014).

58. Deighton, C.M., Walker, D.J., Griffiths, I.D. & Roberts, D.F. The contribution of HLA to rheumatoid arthritis. *Clin Genet* **36**, 178-182 (1989).

59. Begovich, A.B., *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* **75**, 330-337 (2004).

60. Stanford, S.M. & Bottini, N. PTPN22: the archetypal non-HLA autoimmunity gene. *Nat Rev Rheumatol* **10**, 602-611 (2014).

61. Suzuki, A., *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat Genet* **34**, 395-402 (2003).

62. Plenge, R.M., *et al.* Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am J Hum Genet* **77**, 1044-1060 (2005).

63. Kurreeman, F.A., *et al.* A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis. *PLoS Med* **4**, e278 (2007).

64. Kochi, Y., *et al.* A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat Genet* **37**, 478-485 (2005).

65. Viatte, S., Plant, D. & Raychaudhuri, S. Genetics and epigenetics of rheumatoid arthritis. *Nat Rev Rheumatol* **9**, 141-153 (2013).

66. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536-539 (1996).

67. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).

68. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* **14**, 661-673 (2013).

69. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

70. Plenge, R.M., *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* **39**, 1477-1482 (2007).

71. Plenge, R.M., *et al.* TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med* **357**, 1199-1209 (2007).

72. Remmers, E.F., *et al.* STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med* **357**, 977-986 (2007).

73. Barton, A., *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat Genet* **40**, 1156-1159 (2008).

74. Raychaudhuri, S., *et al.* Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat Genet* **41**, 1313-1318 (2009).

75. Stahl, E.A., *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* **42**, 508-514 (2010).

76. Freudenberg, J., *et al.* Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum* **63**, 884-893 (2011).

77. Eyre, S., *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* **44**, 1336-1340 (2012).

78. Okada, Y., *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat Genet* **44**, 511-516 (2012).

79. Okada, Y., *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381 (2014).

80. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).

81. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-266 (2008).

82. Lawrence, J.S. The epidemiology and genetics of rheumatoid arthritis. *Rheumatology* **2**, 1-36 (1969).

83. Aho, K., Koskenvuo, M., Tuominen, J. & Kaprio, J. Occurrence of rheumatoid arthritis in a nationwide series of twins. *J Rheumatol* **13**, 899-902 (1986).

84. Silman, A.J., *et al.* Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br J Rheumatol* **32**, 903-907 (1993).

85. Hensvold, A.H., *et al.* Environmental and genetic factors in the development of anticitrullinated protein antibodies (ACPAs) and ACPA-positive rheumatoid arthritis: an epidemiological investigation in twins. *Ann Rheum Dis* **74**, 375-380 (2015).

86. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**, 1011-1021 (2012).

87. Stahl, E.A., *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44**, 483-489 (2012).

88. Barrera, P., Radstake, T.R., Albers, J.M., van Riel, P.L. & van de Putte, L.B. Familial aggregation of rheumatoid arthritis in The Netherlands: a cross-sectional hospital-based survey. European Consortium on Rheumatoid Arthritis families (ECRAF). *Rheumatology (Oxford)* **38**, 415-422 (1999).

89. Lynn, A.H., Kwoh, C.K., Venglish, C.M., Aston, C.E. & Chakravarti, A. Genetic epidemiology of rheumatoid arthritis. *Am J Hum Genet* **57**, 150-159 (1995).

90. Deighton, C.M. & Walker, D.J. What factors distinguish probands from multicase rheumatoid arthritis same sex sibships from sporadic disease? *J Rheumatol* **19**, 237-241 (1992).

91. Wolfe, F., Kleinheksel, S.M. & Khan, M.A. Prevalence of familial occurrence in patients with rheumatoid arthritis. *Br J Rheumatol* **27 Suppl 2**, 150-152 (1988).

92. Thomas, D.J., Young, A., Gorsuch, A.N., Bottazzo, G.F. & Cudworth, A.G. Evidence for an association between rheumatoid arthritis and autoimmune endocrine disease. *Ann Rheum Dis* **42**, 297-300 (1983).

93. Hemminki, K., Li, X., Sundquist, J. & Sundquist, K. Familial associations of rheumatoid arthritis with autoimmune diseases and related conditions. *Arthritis Rheum* **60**, 661-668 (2009).

94. Grant, S.F., *et al.* The inheritance of rheumatoid arthritis in Iceland. *Arthritis Rheum* **44**, 2247-2254 (2001).

95. Koumantaki, Y., *et al.* Family history as a risk factor for rheumatoid arthritis: a case-control study. *J Rheumatol* **24**, 1522-1526 (1997).

96. Jones, M.A., Silman, A.J., Whiting, S., Barrett, E.M. & Symmons, D.P. Occurrence of rheumatoid arthritis is not increased in the first degree relatives of a population based inception cohort of inflammatory polyarthritis. *Ann Rheum Dis* **55**, 89-93 (1996).

97. del Junco, D., Luthra, H.S., Annegers, J.F., Worthington, J.W. & Kurland, L.T. The familial aggregation of rheumatoid arthritis and its relationship to the HLA-DR4 association. *Am J Epidemiol* **119**, 813-829 (1984).

98. Bowes, J. & Barton, A. Recent advances in the genetics of RA susceptibility. *Rheumatology (Oxford)* **47**, 399-402 (2008).

99. Kurko, J., *et al.* Genetics of rheumatoid arthritis - a comprehensive review. *Clin Rev Allergy Immunol* **45**, 170-179 (2013).

100. de Vries, R. Genetics of rheumatoid arthritis: time for a change! *Curr Opin Rheumatol* **23**, 227-232 (2011).

101. Raychaudhuri, S., *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* **40**, 1216-1223 (2008).

102. Demoruelle, M.K., Deane, K.D. & Holers, V.M. When and where does inflammation begin in rheumatoid arthritis? *Curr Opin Rheumatol* **26**, 64-71 (2014).

103. Yahya, A., *et al.* Smoking is associated with an increased risk of developing ACPA-positive but not ACPA-negative rheumatoid arthritis in Asian populations: evidence from the Malaysian MyEIRA case-control study. *Mod Rheumatol* **22**, 524-531 (2012).

104. Karlson, E.W., *et al.* A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals. *Arthritis Rheum* **42**, 910-917 (1999).

105. Sugiyama, D., *et al.* Impact of smoking as a risk factor for developing rheumatoid arthritis: a meta-analysis of observational studies. *Ann Rheum Dis* **69**, 70-81 (2010).

106. Di Giuseppe, D., Discacciati, A., Orsini, N. & Wolk, A. Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. *Arthritis Res Ther* **16**, R61 (2014).

107. Sopori, M. Effects of cigarette smoke on the immune system. *Nat Rev Immunol* **2**, 372-377 (2002).

108. Wang, H., *et al.* Nicotinic acetylcholine receptor alpha7 subunit is an essential regulator of inflammation. *Nature* **421**, 384-388 (2003).

109. Stolt, P., *et al.* Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann Rheum Dis* **64**, 582-586 (2005).

110. Stolt, P., *et al.* Silica exposure among male current smokers is associated with a high risk of developing ACPA-positive rheumatoid arthritis. *Ann Rheum Dis* **69**, 1072-1076 (2010).

111. Yahya, A., *et al.* Silica exposure is associated with an increased risk of developing ACPA-positive rheumatoid arthritis in an Asian population: evidence from the Malaysian MyEIRA case-control study. *Mod Rheumatol* **24**, 271-274 (2014).

112. Barragan-Martinez, C., *et al.* Organic solvents as risk factor for autoimmune diseases: a systematic review and meta-analysis. *PLoS One* **7**, e51506 (2012).

113. Hart, J.E., *et al.* Ambient air pollution exposures and risk of rheumatoid arthritis: results from the Swedish EIRA case-control study. *Ann Rheum Dis* **72**, 888-894 (2013).

114. Sverdrup, B., *et al.* Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case-control study. *Arthritis Res Ther* **7**, R1296-1303 (2005).

115. Lu, B., Solomon, D.H., Costenbader, K.H. & Karlson, E.W. Alcohol consumption and risk of incident rheumatoid arthritis in women: a prospective study. *Arthritis Rheumatol* **66**, 1998-2005 (2014).

116. Di Giuseppe, D., Alfredsson, L., Bottai, M., Askling, J. & Wolk, A. Long term alcohol intake and risk of rheumatoid arthritis in women: a population based cohort study. *BMJ* **345**, e4230 (2012).

117. Maxwell, J.R., Gowers, I.R., Moore, D.J. & Wilson, A.G. Alcohol consumption is inversely associated with risk and severity of rheumatoid arthritis. *Rheumatology (Oxford)* **49**, 2140-2146 (2010).

118. Kallberg, H., *et al.* Alcohol consumption is associated with decreased risk of rheumatoid arthritis: results from two Scandinavian case-control studies. *Ann Rheum Dis* **68**, 222-227 (2009).

119. Pedersen, M., *et al.* Environmental risk factors differ between rheumatoid arthritis with and without auto-antibodies against cyclic citrullinated peptides. *Arthritis Res Ther* **8**, R133 (2006).

120. Jonsson, I.M., *et al.* Ethanol prevents development of destructive arthritis. *Proc Natl Acad Sci U S A* **104**, 258-263 (2007).

121. Caplan, L., *et al.* Body mass index and the rheumatoid arthritis swollen joint count: an observational study. *Arthritis Care Res (Hoboken)* **65**, 101-106 (2013).

122. Lu, B., *et al.* Being overweight or obese and risk of developing rheumatoid arthritis among women: a prospective cohort study. *Ann Rheum Dis* (2014).

123. Wesley, A., *et al.* Association between body mass index and anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis: results from a population-based case-control study. *Arthritis Care Res (Hoboken)* **65**, 107-112 (2013).

124. James, M., Proudman, S. & Cleland, L. Fish oil and rheumatoid arthritis: past, present and future. *Proc Nutr Soc* **69**, 316-323 (2010).

125. Rosell, M., Wesley, A.M., Rydin, K., Klareskog, L. & Alfredsson, L. Dietary fish and fish oil and the risk of rheumatoid arthritis. *Epidemiology* **20**, 896-901 (2009).

126. Shapiro, J.A., *et al.* Diet and rheumatoid arthritis in women: a possible protective effect of fish consumption. *Epidemiology* **7**, 256-263 (1996).

127. Bengtsson, C., Nordmark, B., Klareskog, L., Lundberg, I. & Alfredsson, L. Socioeconomic status and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann Rheum Dis* **64**, 1588-1594 (2005).

128. Mackie, S.L., *et al.* Relationship between area-level socio-economic deprivation and autoantibody status in patients with rheumatoid arthritis: multicentre cross-sectional study. *Ann Rheum Dis* **71**, 1640-1645 (2012).

129. Knol, M.J., Egger, M., Scott, P., Geerlings, M.I. & Vandenbroucke, J.P. When one depends on the other: reporting of interaction in case-control and cohort studies. *Epidemiology* **20**, 161-166 (2009).

130. Ahlbom, A. & Alfredsson, L. Interaction: A word with two meanings creates confusion. *Eur J Epidemiol* **20**, 563-564 (2005).

131. Rothman, K.J., Greenland, S. & Walker, A.M. Concepts of interaction. *Am J Epidemiol* **112**, 467-470 (1980).

132. Rothman, K.J. Causes. *Am J Epidemiol* **104**, 587-592 (1976).

133. VanderWeele, T.J. Sufficient cause interactions and statistical interactions. *Epidemiology* **20**, 6-13 (2009).

134. Andersson, T., Alfredsson, L., Kallberg, H., Zdravkovic, S. & Ahlbom, A. Calculating measures of biological interaction. *Eur J Epidemiol* **20**, 575-579 (2005).

135. Karlson, E.W., *et al.* Gene-environment interaction between HLA-DRB1 shared epitope and heavy cigarette smoking in predicting incident rheumatoid arthritis. *Ann Rheum Dis* **69**, 54-60 (2010).

136. Morgan, A.W., *et al.* Reevaluation of the interaction between HLA-DRB1 shared epitope alleles, PTPN22, and smoking in determining susceptibility to autoantibody-positive and autoantibody-negative rheumatoid arthritis in a large UK Caucasian population. *Arthritis Rheum* **60**, 2565-2576 (2009).

137. Mahdi, H., *et al.* Specific interaction between genotype, smoking and autoimmunity to citrullinated alpha-enolase in the etiology of rheumatoid arthritis. *Nat Genet* **41**, 1319-1324 (2009).

138. Kallberg, H., *et al.* Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet* **80**, 867-875 (2007).

139. Makrygiannakis, D., *et al.* Smoking increases peptidylarginine deiminase 2 enzyme expression in human lungs and increases citrullination in BAL cells. *Ann Rheum Dis* **67**, 1488-1492 (2008).

140. Mikuls, T.R., *et al.* Impact of interactions of cigarette smoking with NAT2 polymorphisms on rheumatoid arthritis risk in African Americans. *Arthritis Rheum* **64**, 655-664 (2012).

141. Keenan, B.T., *et al.* Effect of interactions of glutathione S-transferase T1, M1, and P1 and HMOX1 gene promoter polymorphisms with heavy smoking on the risk of rheumatoid arthritis. *Arthritis Rheum* **62**, 3196-3210 (2010).

142. Costenbader, K.H., Chang, S.C., De Vivo, I., Plenge, R. & Karlson, E.W. Genetic polymorphisms in PTPN22, PADI-4, and CTLA-4 and risk for rheumatoid arthritis in two longitudinal cohort studies: evidence of gene-environment interactions with heavy cigarette smoking. *Arthritis Res Ther* **10**, R52 (2008).

143. Rantapaa-Dahlqvist, S., *et al.* Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis Rheum* **48**, 2741-2749 (2003).

144. Ekbom, A. The Swedish Multi-generation Register. *Methods Mol Biol* **675**, 215-220 (2011).

145. Ludvigsson, J.F., *et al.* External review and validation of the Swedish national inpatient register. *BMC Public Health* **11**, 450 (2011).

146. Baecklund, E., *et al.* Association of chronic inflammation, not its treatment, with increased lymphoma risk in rheumatoid arthritis. *Arthritis Rheum* **54**, 692-701 (2006).

147. Ding, B., Kallberg, H., Klareskog, L., Padyukov, L. & Alfredsson, L. GEIRA: gene-environment and gene-gene interaction research application. *Eur J Epidemiol* **26**, 557-561 (2011).

148. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).

149. Iles, M.M. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet* **4**, e33 (2008).

150. McCarthy, M.I., *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369 (2008).

151. Rothman, K.J. *Epidemiology : an introduction*, (Oxford University Press, New York, NY, 2012).

152. Little, R.J., *et al.* The prevention and treatment of missing data in clinical trials. *N Engl J Med* **367**, 1355-1360 (2012).

153. Donders, A.R., van der Heijden, G.J., Stijnen, T. & Moons, K.G. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* **59**, 1087-1091 (2006).

154. Sterne, J.A., *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393 (2009).

155. He, Y. Missing data analysis using multiple imputation: getting to the heart of the matter. *Circ Cardiovasc Qual Outcomes* **3**, 98-105 (2010).

156. Perricone, C., Ceccarelli, F. & Valesini, G. An overview on the genetic of rheumatoid arthritis: a never-ending story. *Autoimmun Rev* **10**, 599-608 (2011).

157. Weyand, C.M. & Goronzy, J.J. Association of MHC and rheumatoid arthritis. HLA polymorphisms in phenotypic variants of rheumatoid arthritis. *Arthritis Res* **2**, 212-216 (2000).

158. van der Helm-van Mil, A.H., *et al.* The HLA-DRB1 shared epitope alleles differ in the interaction with smoking and predisposition to antibodies to cyclic citrullinated peptide. *Arthritis Rheum* **56**, 425-432 (2007).

159. Laki, J., *et al.* Very high levels of anti-citrullinated protein antibodies are associated with HLA-DRB1*15 non-shared epitope allele in patients with rheumatoid arthritis. *Arthritis Rheum* **64**, 2078-2084 (2012).

160. van der Woude, D., *et al.* Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: a meta-analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in four European populations. *Arthritis Rheum* **62**, 1236-1245 (2010).

161. Han, B., *et al.* Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am J Hum Genet* **94**, 522-532 (2014).

162. Sofat, N. & Keat, A. Alcohol intake in rheumatic disease: good or bad? *Rheumatology (Oxford)* **41**, 125-128 (2002).

163. Zou, G.Y. On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol* **168**, 212-224 (2008).

164. du Montcel, S.T., *et al.* New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis Rheum* **52**, 1063-1068 (2005).

165. Michou, L., *et al.* Validation of the reshaped shared epitope HLA-DRB1 classification in rheumatoid arthritis. *Arthritis Res Ther* **8**, R79 (2006).

166. Barnetche, T., Constantin, A., Cantagrel, A., Cambon-Thomsen, A. & Gourraud, P.A. New classification of HLA-DRB1 alleles in rheumatoid arthritis susceptibility: a combined analysis of worldwide samples. *Arthritis Res Ther* **10**, R26 (2008).

167. Liu, Y., *et al*. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**, 142-147 (2013).